

Департамент образования и науки города Москвы Государственное автономное образовательное учреждение высшего образования города Москвы «Московский городской педагогический университет» Институт цифрового образования Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

Инструменты для хранения и обработки больших данных

Практическая работа 01-1

Тема:

Визуализация данных из CSV-файла в DataLeans

Выполнила: Соколова М. С., группа: АДЭУ-201

Преподаватель: Босенко Т. М.

Москва

2023

Данные, которые были предоставлены со станицы [Moscow tutors | Kaggle](#):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	Categories	Price	Score	Format	Reviews_number	Experience	Status	Location	Tags	Audience	Video_presentation	Photo												
2	[Немецкий язык, Испанский язык]	1800.5.0	[дистанционно]		26.0	21.0	Частный преподаватель																	
3	[Математика]	2500.4.9	[у репетитора, дистанционно]		41.0	29.0	Частный преподаватель	[м. Шуйкинская]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'школьный курс', 'Алгебра', 'Геометрия', 'Тригонометрия']	[Школьники 4-11 классов, 'Студенты']														
4	[Английский язык]	1500.5.0	[у репетитора, у ученика, дистанционно]		14.0	11.0	Частный преподаватель	[м. Чертановская]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'ФСЕ', 'KET', 'PET', 'YLE', '+1', 'Общий курс']	[Дети 6-7 лет, 'Школьные 1-11 классов']														
5	[Химия]	1300.5.0	[дистанционно]		41.0	39.0	Частный преподаватель																	
6	[Математика]	1500.5.0	[у репетитора, у ученика, дистанционно]		35.0	9.0	Школьный преподаватель	[м. Отрадное, 'м. Бабушкинская', 'м. Свиблово', 'м. Бибирево']	[ГОЭ (ГИА), 'ЕГЭ', 'школьный курс', 'Алгебра', 'Геометрия']	[Школьники 4-11 классов, 'Студенты']														
7	[Математика, 'Физика']	2000.4.9	[у репетитора, дистанционно]		33.0	9.0	Аспирант	[м. Новогиреево, 'г. Балашиха']	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'школьный курс', 'Алгебра', 'Аналитическая геометрия', 'Высшая математика']	[Школьники 4-11 классов, 'Студенты']														
8	[Английский язык]	2000.4.9	[у репетитора, у ученика, дистанционно]		12.0	9.0	Частный преподаватель	[м. Люблино]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'Бизнес-курс', 'CAE', 'FCE', 'IELTS', '+6', 'KET', 'PET', 'TOEFL', 'Общий курс']	[Школьники 4-11 классов, 'Студенты']														
9	[Английский язык]	6000.5.0	[дистанционно]		23.0	17.0	Носитель языка																	
10	[Английский язык]	2000.5.0	[у репетитора, у ученика, дистанционно]		26.0	13.0	Частный преподаватель	[м. Некрасовка]	[ГОЭ (ГИА), 'ЕГЭ', 'Бизнес-курс', 'Общий курс']	[Дети 4-5 лет, 'Дети 6-7 лет', 'Школьники 1-11 классов', 'Студенты']														
11	[Русский язык, 'Литература', 'Обществознание', 'История']	3000.5.0	[у репетитора, дистанционно]		64.0	26.0	Преподаватель вуза	[м. Марксистская]	[ГОЭ (ГИА), 'ЕГЭ', 'школьный курс']	[Школьники 4-11 классов]	No, Yes													
12	[Математика]	2000.4.9	[у ученика, дистанционно]		12.0	23.0	Частный преподаватель	[м. Бульвар Рокоссовского, 'м. Преображенская площадь', 'м. Семёновская']	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'школьный курс', 'Алгебра', 'Геометрия']	[Школьники 4-11 классов, 'Студенты']														
13	[Музыка]	2000.4.9	[у репетитора, у ученика, дистанционно]		14.0	6.0	Частный преподаватель	[м. Коломенская]	[подготовка к поступлению в муз. уч. заведения, 'Акустическая гитара', 'Аранжировка', 'Бас-гитара', 'Вокал', 'Гаваяская гитара']	[Школьники 4-11 классов, 'Студенты']														
14	[Физика]	2000.5.0	[у репетитора, у ученика, дистанционно]		21.0	36.0	Частный преподаватель	[м. Молодёжная]	[ГОЭ (ГИА), 'ЕГЭ', 'школьный курс']	[Школьники 7-11 классов, 'Взрослые']	No, Yes													
15	[История, 'Обществознание']	1350.4.9	[у репетитора, у ученика, дистанционно]		39.0	20.0	Частный преподаватель	[м. Рязанский проспект]	[ГОЭ (ГИА), 'ЕГЭ', 'школьный курс', 'Вузовский курс']	[Школьники 6-11 классов, 'Студенты']	Yes, Yes													
16	[Английский язык]	1400.4.9	[у репетитора, дистанционно]		39.0	20.0	Частный преподаватель	[м. Провальский]	[ГОЭ (ГИА), 'ЕГЭ', 'Бизнес-курс', 'Общий курс']	[Школьники 1-11 классов, 'Студенты', 'Взрослые']	Yes, Yes													
17	[Химия]	1400.5.0	[у репетитора, у ученика, дистанционно]		42.0	13.0	Частный преподаватель	[г. Железнодорожный]	[ГОЭ (ГИА), 'ЕГЭ', 'школьный курс', 'Вузовский курс']	[Школьники 8-11 классов, 'Студенты']	No, Yes													
18	[Английский язык]	1000.4.8	[дистанционно]		9.0	7.0	Частный преподаватель																	
19	[Английский язык]	4000.5.0	[у репетитора, дистанционно]		13.0	9.0	Преподаватель вуза	[м. Молодёжная]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'Бизнес-курс', 'CAE', 'CPE', 'FCE', '+5', 'KET', 'PET', 'Общий курс', 'Американский английский']	[Школьники 4-11 классов, 'Студенты']														
20	[Русский как иностранный, 'История', 'Обществознание', 'География', 'Другой', 'Биология']	2000.5.0	[у ученика, дистанционно]		33.0	16.0	Преподаватель вуза	[м. Селигерская]	[Язык-посредник: английский]	[Школьники 5-11 классов]														
21	[Японский язык, 'Русский как иностранный']	1400.5.0	[дистанционно]		41.0	14.0	Частный преподаватель																	
22	[Английский язык]	1900.4.9	[у репетитора, у ученика, дистанционно]		55.0	14.0	Частный преподаватель	[м. Отрадное]	[ГОЭ (ГИА), 'ЕГЭ', 'BEC', 'CAE', 'FCE', 'IELTS', 'KET', '+4', 'PET', 'TOEFL', 'Общий курс', 'Британский английский']	[Школьники 4-11 классов, 'Студенты']														
23	[Английский язык]	2000.5.0	[дистанционно]		48.0	9.0	Частный преподаватель																	
24	[Английский язык, 'Немецкий язык', 'Русский язык']	1500.4.9	[у репетитора, у ученика, дистанционно]		83.0	11.0	Частный преподаватель	[м. Академическая]	[ГОЭ (ГИА), 'ЕГЭ', 'FCE', 'GMAT', 'IELTS', 'PET', 'TOEFL', '+1', 'Общий курс']	[Школьники 4-11 классов, 'Студенты']														
25	[Обществознание, 'Другой', 'Музыка', 'Редкие иностранные языки']	2000.5.0	[у репетитора, у ученика, дистанционно]		21.0	9.0	Частный преподаватель	[м. Медведково, 'м. Митино', 'м. Выхино']	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'школьный курс', 'Алгебра', 'Геометрия', 'Физика', 'Химия', 'Биология', 'История', 'Литература', 'Обществознание']	[Школьники 4-11 классов, 'Студенты']														
26	[Русский язык, 'Математика']	1400.5.0	[у ученика, дистанционно]		13.0	8.0	Частный преподаватель	[г. Жулебино, 'г. Люберцы']	[ГОЭ (ГИА), 'ЕГЭ', 'школьный курс']	[Школьники 5-11 классов]	Yes, Yes													
27	[Математика, 'Начальная школа']	2000.4.8	[у репетитора, у ученика, дистанционно]		73.0	13.0	Частный преподаватель	[г. Павловский Посад]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'школьный курс', 'Алгебра', 'Алгебра логики', 'Геометрия', 'Физика', 'Химия', 'Биология', 'История', 'Литература', 'Обществознание']	[Школьники 4-11 классов, 'Студенты']														
28	[Русский язык, 'Литература']	1500.5.0	[у репетитора, у ученика, дистанционно]		40.0	22.0	Частный преподаватель	[м. Ростокино]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'школьный курс', 'Вузовский курс']	[Дети 6-7 лет, 'Школьники 4-11 классов, 'Студенты']														

Переходим в Colab Research [Big-Data.Pr 01-1 - Colaboratory \(google.com\)](#):

Загрузка

Ссылка на данные: <https://drive.google.com/file/d/15MWNyAD2xrqvWCFmYsYxILcZTPUYD/view?usp=sharing>

Ссылка на Kaggle: <https://www.kaggle.com/datasets/vadimantipov/moscow-tutors?resource=download>

Ссылка на дашборд DataLens: <https://datalens.yandex.cc/jd10vq5iw62>

1 #Загружаем файл с Google Диска

2 ! gdown --id 15MWNyAD2xrqvWCFmYsYxILcZTPUYD

/usr/local/lib/python3.8/dist-packages/gdown/cli.py:127: FutureWarning: Option '--id' was deprecated in version 4.3.1 and will be removed in 5.0. You don't need to pass it anymore to use a file ID.

warnings.warn(DeprecationWarning)

Downloading...

From: <https://drive.google.com/uc?id=15MWNyAD2xrqvWCFmYsYxILcZTPUYD>

To: /content/Tech.zip

100% 1.68M/1.68M [00:00<00:00, 116MB/s]

[2] 1 #Разархивируем загруженный zip-файл с Google Диска

2 ! unzip Tech.zip

Archive: Tech.zip

inflating: tutors_rus_2021_10_06/tutors_rus_2021_10_06.csv

1 import pandas as pd

2 import numpy as np

3 import seaborn as sns

4 import matplotlib.pyplot as plt

5 from ast import literal_eval

1 #Записываем csv-файл в датафрейм и смотрим первые 5 строк

2 df = pd.read_csv('/content/tutors_rus_2021_10_06/tutors_rus_2021_10_06.csv')

3 df.head()

	Categories	Price	Score	Format	Reviews_number	Experience	Status	Location	Tags	Audience	Video_presentation	Photo
0	[Немецкий язык, Испанский язык]	1800	5.0	[дистанционно]	26.0	21.0	Частный преподаватель	NaN	NaN	NaN	NaN	NaN
1	[Математика]	2500	4.9	[у репетитора, дистанционно]	41.0	29.0	Частный преподаватель	[м. Шуйкинская]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'школьный курс', 'Алгебра', 'Геометрия', 'Тригонометрия']	[Школьники 4-11 классов, 'Студенты']	No	Yes
2	[Английский язык]	1500	5.0	[у репетитора, у ученика, дистанционно]	14.0	11.0	Частный преподаватель	[м. Чертановская]	[ГОЭ (ГИА), 'ЕГЭ', 'подготовка к олимпиадам', 'ФСЕ', 'KET', 'PET', 'YLE', '+1', 'Общий курс']	[Дети 6-7 лет, 'Школьные 1-11 классов', 'Студенты']	Yes	Yes
3	[Химия]	1300	5.0	[дистанционно]	41.0	39.0	Частный преподаватель	NaN	NaN	NaN	NaN	NaN
4	[Математика]	1500	5.0	[у репетитора, у ученика, дистанционно]	35.0	9.0	Школьный преподаватель	[м. Отрадное, 'м. Бабушкинская', 'м. Свиблово', 'м. Бибирево']	[ГОЭ (ГИА), 'ЕГЭ', 'школьный курс', 'Алгебра', 'Геометрия']	[Школьники 5-11 классов]	No	Yes

1 #Смотрим размерность датафрейма

2 df.shape

(88699, 12)

Обработка данных

1 #Смотрим сколько процентов данных с пустыми значениями

2 df.isna().mean()

Categories	0.000000
Price	0.000000
Score	0.000000
Format	0.000000
Reviews_number	0.712229
Experience	0.712229
Status	0.712229
Location	0.737393
Tags	0.737539
Audience	0.737551
Video_presentation	0.737551
Photo	0.737551
dtype:	float64

Почти 74 % данных отсутствуют или пусты для 8 последних столбцов.

Учитывая тот факт, что набор данных большой и недостающие значения распределены по всему набору данных равномерно, мы можем удалить строки, содержащие любое количество значений NaN, и по-прежнему иметь репрезентативный большой набор данных

1 #Удаляем строки, в которых есть ячейки со значением NaN и восстанавливаем индексацию

2 df = df.dropna(axis=0).reset_index(drop=True)

3 df.tail()

	Categories	Price	Score	Format	Reviews_number	Experience	Status	Location	Tags	Audience	Video_presentation	Photo
23274	[Русский язык, 'Занятия с дошкольниками]	1300	3.6	[у ученика]	1.0	18.0	Частный преподаватель	[м. Алтуфьево]	[Школьный курс]	[Школьники 1-11 классов]		No Yes
23275	[Занятия с дошкольниками, Начальная школа]	1000	4.6	[у ученика]	5.0	14.0	Частный преподаватель	[м. Щёлковская]	[английский для малышей, 'Общий курс', 'Мате...	[Дети 4-5 лет', 'Дети 6-7 лет]		No No
23276	[Английский язык, 'Русский как иностранный,...	700	4.7	[у репетитора, у ученика, 'дистанционно]	6.0	13.0	Частный преподаватель	[м. Университет]	[бизнес-курс, 'Общий курс]	[Дети 1-3 года', 'Дети 4-5 лет', 'Дети 6-7 ле...		No No
23277	[Французский язык]	1000	5.0	[у ученика]	1.0	38.0	Школьный преподаватель	[г. Люберцы]	[ОГЭ ('ГИА), 'ЕГЭ, 'бизнес-курс, 'Общий курс]	[Школьники 1-11 классов, 'Студенты', 'Варосл...		No No
23278	[Английский язык]	500	5.0	[у репетитора, у ученика, 'дистанционно]	1.0	4.0	Студент	[м. Курская (радиальная)', 'м. Курская (кольц...	[ОГЭ ('ГИА), 'A-Level, 'Общий курс]	[Дети 6-7 лет', 'Школьники 1-9 классов]		No No

1 #Просматриваем типы данных

2 df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23279 entries, 0 to 23278
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   Categories              23279 non-null object
1   Price                  23279 non-null int64  
2   Score                  23279 non-null float64
3   Format                  23279 non-null object
4   Reviews_number         23279 non-null float64
5   Experience              23279 non-null float64
6   Status                  23279 non-null object
7   Location                23279 non-null object
8   Tags                    23279 non-null object
9   Audience                23279 non-null object
10  Video_presentation      23279 non-null object
11  Photo                   23279 non-null object
dtypes: float64(3), int64(1), object(8)
memory usage: 2.1+ MB
```

- Описание полей:
1. **Categories** - список преподаваемых предметов;
 2. **Price** - Цена в рублях за час;
 3. **Score** - Средний балл, основанный на отзывах;
 4. **Format** - Варианты форматов обучения: дистанционно, у репетитора, у студента;
 5. **Reviews_number** - Количество отзывов в профиле преподавателя;
 6. **Experience** - Опыт в годах;
 7. **Status** - Текущий статус преподавателя: Частный репетитор, Школьный учитель, Аспирант, носитель языка, профессор университета, Студент, не указано;
 8. **Местоположение** - Станции метро или города Московской области;
 9. **Tags** - Услуги репетитора. Они изложены преподавателями и могут отличаться;
 10. **Audience** - целевая аудитория преподавателя. Например: студенты, учащиеся 10 классов и т.д;
 11. **Video_presentation** - Доступность видеопрезентации;
 12. **Photo** - Наличие фотографий в профиле.

Столбцы Reviews_number и Experience содержат числа с плавающей запятой, хотя должны быть целыми числами.

1 #Изменение типов данных

2 df['Reviews_number'] = df['Reviews_number'].astype(int)

3 df['Experience'] = df['Experience'].astype(int)

Изменяем столбец Format

1 df['Format'].unique()

```
array(['у репетитора', 'дистанционно'],
      ['у репетитора', 'у ученика', 'дистанционно'],
      ['у ученика', 'дистанционно'], ['у репетитора'],
      ['у репетитора', 'у ученика'], ['у ученика']], dtype=object)
```

1 def format(x):

2 if x == ["у репетитора", 'дистанционно']:

3 return 'У репетитора или дистанционно'

4 if x == ["у репетитора", 'у ученика', 'дистанционно']:

5 return 'Любой формат'

6 if x == ["у ученика", 'дистанционно']:

7 return 'У ученика или дистанционно'

8 if x == ["у ученика"]:

9 return 'У ученика'

10 if x == ["у репетитора"]:

11 return 'У репетитора'

12 if x == ["у репетитора", 'у ученика']:

13 return 'У репетитора или ученика'

1 df['Format_new'] = df['Format'].apply(format)

Разбиваем столбец Price на три сегмента:

1. Низкий ценовой сегмент
2. Средний ценовой сегмент
3. Высокий ценовой сегмент

1 df['Price_group_count'] = pd.qcut(df['Price'], 3)

Удаляем столбец Tags, т.к он не пригодиться нам для анализа

1 df = df.drop('Tags', axis=1)

✓

[100]

1 df.head()

0

OK

	Categories	Price	Score	Format	Reviews_number	Experience	Status	Location	Audience	Video_presentation	Photo	Format_new	Price_group_count	
0	[Математика]	2500	4.9	[у репетитора, дистанционно]	41	29	Частный преподаватель	[м. Щукинская]	[Школьники 4-11 классов, Студенты]		No	Yes	У репетитора или дистанционно	(1100.0, 6700.0]
1	[Английский язык]	1500	5.0	[у репетитора, ученика, дистанционно]	14	11	Частный преподаватель	[м. Чертановская]	[Дети 6-7 лет, Школьники 1-11 классов, Ст...		Yes	Yes	Любой формат	(1100.0, 6700.0]
2	[Математика]	1500	5.0	[у репетитора, ученика, дистанционно]	35	9	Школьный преподаватель	[м. Отрадное, м. Бабушкинская, м. Свиблово...	[Школьники 5-11 классов]		No	Yes	Любой формат	(1100.0, 6700.0]
3	[Математика, Физика]	2000	4.9	[у репетитора, дистанционно]	33	9	Аспирант	[м. Новогиреево, г. Балашиха]	[Школьники 7-11 классов, Студенты, Взросл...		No	Yes	У репетитора или дистанционно	(1100.0, 6700.0]
4	[Английский язык]	2000	4.9	[у репетитора, ученика, дистанционно]	12	9	Частный преподаватель	[м. Люблино]	[Дети 1-3 года, Дети 4-5 лет, Дети 6-7 ле...		Yes	Yes	Любой формат	(1100.0, 6700.0]

▼ Анализ

✓

[20]

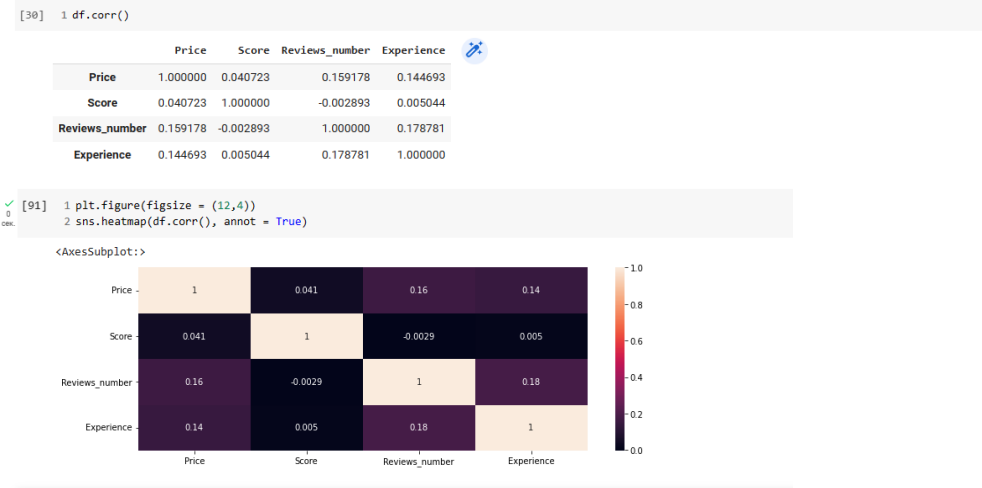
1 #Просматриваем статистику
2 df.describe()

0

OK

	Price	Score	Reviews_number	Experience
count	23279.000000	23279.000000	23279.000000	23279.000000
mean	1075.114910	4.892109	6.214614	15.244340
std	511.798633	0.229869	9.177704	9.929735
min	150.000000	2.200000	0.000000	0.000000
25%	700.000000	4.900000	1.000000	8.000000
50%	1000.000000	5.000000	3.000000	13.000000
75%	1300.000000	5.000000	7.000000	20.000000
max	6700.000000	5.000000	172.000000	64.000000

▼ Корреляция Пирсона



▼ Корреляция Фика



Группировка

```
[73] 1 def group(column):
2     #делаем группировку по столбцу , добавляя среднее значение показателей
3     group_df = df.groupby(column, dropna=False)[['Price', 'Score', 'Experience', 'Reviews_number']].agg(['count', 'mean'])
4     #Убираем двухэтажные названия
5     group_df.columns = ['_'.join(col).strip() for col in group_df.columns.values]
6     #Удаляем повторяющиеся столбцы количества
7     group_df = group_df.drop(columns=['Score_count', 'Experience_count', 'Reviews_number_count'])
8     #Переименовываем столбцы
9     group_df = group_df.rename(columns={'Price_count': 'Количество', 'Price_mean': 'Средняя цена', 'Score_mean': 'Средняя оценка',
10     'Experience_mean': 'Средний страх', 'Reviews_number_mean': 'Среднее кол-во отзывов'})
11     return group_df.sort_values('Средняя цена', ascending=False)
12
```

```
[74] 1 group('Price_group_count')
```

	Количество	Средняя цена	Средняя оценка	Средний страх	Среднее кол-во отзывов
Price_group_count					
(1100.0, 6700.0]	7758	1625.264243	4.903442	17.164991	7.955401
(800.0, 1100.0]	7101	990.550627	4.892860	15.462048	6.342487
(149.999, 800.0]	8420	639.536817	4.881033	13.291093	4.502850

Можем увидеть четкую зависимость между всеми метриками и ценовым сегментом. Чем выше стоимость, тем выше все показатели.

```
[25] 1 group('Status')
```

	Количество	Средняя цена	Средняя оценка	Средний страх	Среднее кол-во отзывов
Status					
Носитель языка	438	1282.305936	4.826484	10.678082	5.392694
Преподаватель вуза	1862	1231.364125	4.900859	21.162728	8.264232
Аспирант	167	1132.634731	4.905988	7.365269	5.263473
Частный преподаватель	16013	1081.012303	4.891482	15.420783	6.364766
Школьный преподаватель	2994	1079.325317	4.909920	18.608550	6.395792
Студент	1805	799.030471	4.873740	3.830471	2.755125

Можем увидеть, что максимальная средняя стоимость у носителя языка, а минимальная у студента, это логично.

```
[27] 1 group('Format_new')
```

	Количество	Средняя цена	Средняя оценка	Средний страх	Среднее кол-во отзывов
Format_new					
У ученика	2292	1143.542757	4.880846	15.194154	4.370419
У репетитора или ученика	2720	1138.216912	4.878015	18.522426	5.132353
У репетитора	921	1123.778502	4.899783	23.676439	5.685125
У репетитора или дистанционно	3619	1118.872617	4.914396	17.508151	7.168555
Любой формат	9138	1053.250164	4.892362	14.351499	6.809696
У ученика или дистанционно	4589	1002.800174	4.886468	11.626716	5.946176

Можем увидеть, что максимальная стоимость при формате обучения у ученика, а максимальная стоимость и максимальная оценка у репетитора или дистанционно.

```
[28] 1 group('Photo')
```

	Количество	Средняя цена	Средняя оценка	Средний страх	Среднее кол-во отзывов
Photo					
Yes	20495	1092.698219	4.894682	14.765748	6.404489
No	2784	945.671695	4.873168	18.767601	4.816810

```
[29] 1 group('Video_presentation')
```

	Количество	Средняя цена	Средняя оценка	Средний страх	Среднее кол-во отзывов
Video_presentation					
Yes	562	1451.601423	4.925089	14.153025	12.414591
No	22717	1065.800942	4.891293	15.271339	6.061232

```
[30] 1 photo_video = df.pivot_table(index='Photo', columns='Video_presentation', values='Price', aggfunc=['mean', 'count'])
2 photo_video.columns = ['_'.join(col).strip() for col in photo_video.columns.values]
3 photo_video
```

	mean_No	mean_Yes	count_No	count_Yes
Photo				
No	945.920201	600.000000	2782	2
Yes	1082.530725	1454.642857	19935	560

▼ Отдельные сводные таблицы

```
[54] 1 def change(column):
2     #Удаляем лишние знаки и преобразовываем в список столбец Format
3     for i in [column]:
4         df[i] = df[i].apply(lambda s: list(literal_eval(str(s))))
5
6     #Разделяем каждый список на отдельные строки
7     change_df = df[column].explode()
8
9     #Добавляем сводную таблицу
10    change_df = df.join(pd.crosstab(change_df.index, change_df))
11
12    #Выбираем нужные столбцы и считаем сумму
13    change_dff = change_df.columns[13:13+len(change_df)]
14    df_change = change_df[change_dff].sum()
15    return df_change.sort_values(ascending=False)
```

```
[55] 1 change('Format')
у ученика      18739
дистанционно  17346
у репетитора   16398
dtype: int64
```

```
[70] 1 change('Audience')
Студенты      13192
Взрослые      12251
Школьники 1-11 классов  8281
Дети 6-7 лет   6750
Дети 4-5 лет   4380
...
Школьники 4-5      1
Школьники 2        1
Школьники 5        1
Школьники 10 класса  1
Школьники 5-6      1
Length: 107, dtype: int64
```

```
[34] 1 change('Categories').head(10)
Английский язык      7484
Математика           5800
Русский язык         3331
Физика               2112
Музыка               2026
Начальная школа      1756
Другой               1580
Химия                1473
Занятия с дошкольниками 1432
Обществознание       1393
dtype: int64
```

```
[81] 1 change('Location').head(10)
м. Университет      412
м. Юго-Западная    410
Ленинский район     393
м. Академическая    374
Центральный район   346
Советский район     333
м. Бабушкинская     303
м. Новогиреево      293
м. ВДНХ             285
м. Щёлковская       274
dtype: int64
```

▼ Выгрузка

```
[ ] 1 from google.colab import files
2 df.to_csv('df.csv')
3 files.download('df.csv')
```

```
[ ] 1 from google.colab import files
2 df1.to_csv('df1.csv')
3 files.download('df1.csv')
```

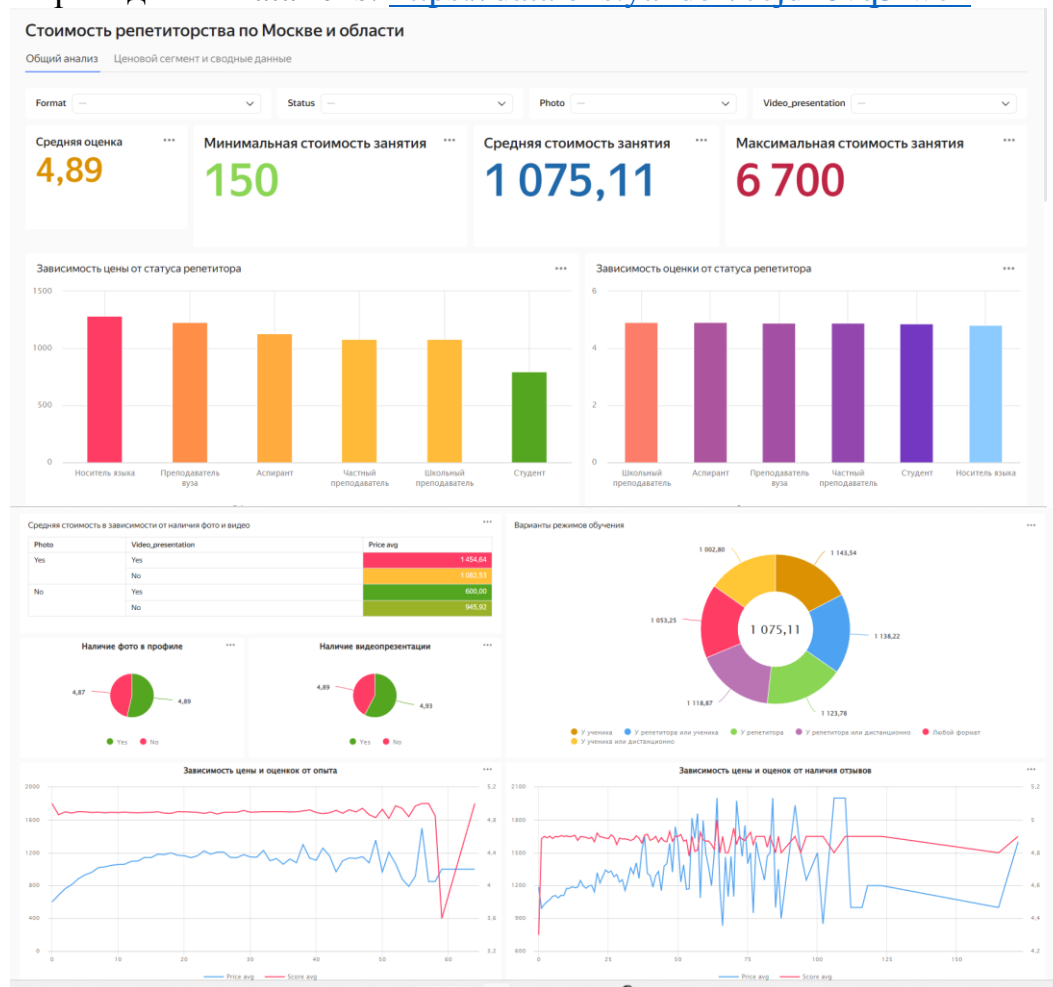
```
[ ] 1 from google.colab import files
2 df2.to_csv('df2.csv')
3 files.download('df2.csv')
```

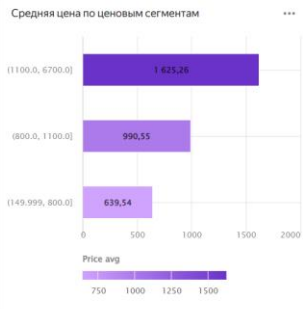
```
▶ 1 from google.colab import files
2 df3.to_csv('df3.csv')
3 files.download('df3.csv')
```

Данные после обработки:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Column1	Categories	Price	Score	Format	Reviews_n	Experience	Status	Location	Audience	Video_pre	Photo	Format_n	Price_group	count	
2	0	['Математ	2500	4.9	['у репети	41	29	Частный г	['м. Щуки	['Школьные	No	Yes	У репети	(1100.0, 6700.0]		
3	1	['Английс	1500	5.0	['у репети	14	11	Частный г	['м. Черта	['Дети 6-7	Yes	Yes	Любой фс	(1100.0, 6700.0]		
4	2	['Математ	1500	5.0	['у репети	35	9	Школьные	['м. Отрад	['Школьные	No	Yes	Любой фс	(1100.0, 6700.0]		
5	3	['Математ	2000	4.9	['у репети	33	9	Аспирант	['м. Новог	['Школьные	No	Yes	У репети	(1100.0, 6700.0]		
6	4	['Английс	2000	4.9	['у репети	12	9	Частный г	['м. Любл	['Дети 1-3	Yes	Yes	Любой фс	(1100.0, 6700.0]		
7	5	['Английс	2000	5.0	['у репети	26	13	Частный г	['м. Некра	['Дети 4-5	Yes	Yes	Любой фс	(1100.0, 6700.0]		
8	6	['Русский	3000	5.0	['у репети	64	26	Преподав	['м. Маркс	['Школьные	No	Yes	У репети	(1100.0, 6700.0]		
9	7	['Математ	2000	4.9	['у ученик	12	23	Частный г	['м. Бульв	['Школьные	Yes	Yes	У ученика	(1100.0, 6700.0]		
10	8	['Музыка]	2000	4.9	['у репети	14	6	Частный г	['м. Колол	['Дети 1-3	Yes	Yes	Любой фс	(1100.0, 6700.0]		
11	9	['Физика]	2000	5.0	['у репети	21	36	Частный г	['м. Моло	['Школьные	No	Yes	Любой фс	(1100.0, 6700.0]		
12	10	['История	1350	4.9	['у репети	32	12	Частный г	['м. Рязан	['Школьные	No	Yes	Любой фс	(1100.0, 6700.0]		
13	11	['Английс	1400	4.9	['у репети	39	20	Частный г	['м. Пражс	['Школьные	Yes	Yes	У репети	(1100.0, 6700.0]		
14	12	['Химия]	1400	5.0	['у репети	42	13	Частный г	['г. Желез	['Школьные	No	Yes	Любой фс	(1100.0, 6700.0]		
15	13	['Английс	4000	5.0	['у репети	13	9	Преподав	['м. Моло	['Дети 6-7	Yes	Yes	У репети	(1100.0, 6700.0]		
16	14	['Русский	2000	5.0	['у ученик	33	16	Преподав	['м. Селиг	['Школьные	No	Yes	У ученика	(1100.0, 6700.0]		
17	15	['Английс	1900	4.9	['у репети	55	14	Частный г	['м. Отрад	['Школьные	No	Yes	Любой фс	(1100.0, 6700.0]		
18	16	['Английс	1500	4.9	['у репети	83	11	Частный г	['м. Акаде	['Школьные	No	Yes	Любой фс	(1100.0, 6700.0]		
19	17	['Обществе	2000	5.0	['у репети	21	9	Частный г	['м. Медв	['Школьные	Yes	Yes	Любой фс	(1100.0, 6700.0]		
20	18	['Русский	1400	5.0	['у ученик	13	8	Частный г	['м. Жуле	['Школьные	Yes	Yes	У ученика	(1100.0, 6700.0]		
21	19	['Математ	2000	4.8	['у ученик	73	13	Частный г	['г. Павло	['Школьные	No	Yes	У ученика	(1100.0, 6700.0]		
22	20	['Русский	1500	5.0	['у репети	40	22	Частный г	['м. Рост	['Дети 6-7	No	Yes	Любой фс	(1100.0, 6700.0]		
23	21	['Информ	1100	4.9	['у репети	82	15	Частный г	['м. Новог	['Студент	No	Yes	Любой фс	(800.0, 1100.0]		
24	22	['Математ	2000	5.0	['у репети	13	6	Частный г	['м. Улиц	['Школьные	Yes	Yes	Любой фс	(1100.0, 6700.0]		
25	23	['Изобраз	1000	5.0	['у репети	63	12	Частный г	['м. Арбат	['Школьные	No	Yes	У репети	(800.0, 1100.0]		
26	24	['Изобраз	600	5.0	['у репети	17	3	Студент	['м. Селиг	['Дети 1-3	Yes	Yes	Любой фс	(149.999, 800.0]		
27	25	['Английс	1500	5.0	['у репети	20	5	Частный г	['м. Дмит	['Дети 1-3	Yes	Yes	Любой фс	(1100.0, 6700.0]		
28	26	['Математ	2000	4.9	['у репети	33	49	Частный г	['м. Алекс	['Школьные	Yes	Yes	У репети	(1100.0, 6700.0]		
29	27	['Русский	2000	4.9	['у репети	60	21	Частный г	['м. Рязан	['Школьные	No	Yes	У ученика	(1100.0, 6700.0]		

Переходим в DataLens: <https://datalens.yandex/ccjd10vq5iwc2>





Сводные данные

