

# Исследовательский проект по НИС «Анализ данных в Python»

Ольга Шавочкина

Магомед Бекаев

# Data Analysis

## Датасет

Нашей командой был выбран набор данных World University Rating, а точнее, Times Higher Education World University Rankings -- *"глобальный рейтинг университетов, ежегодно публикуемый журналом Times Higher Education (THE). Рейтинг THE считается одним из наиболее авторитетных международных рейтингов университетов наряду с Academic Ranking of World Universities и QS World University Rankings"*



# Data Analysis

- world\_rank – место университета в мире
- university\_name – название университета
- country – страна университета
- teaching – оценка образовательной деятельности, от 0 до 100
- international – открытость к другим странам, от 0 до 100
- research – исследовательская деятельность, от 0 до 100
- citations – упоминания, цитирования университета, от 0 до 100
- income – доход, от 0 до 100, 218 отсутствующих значений
- total\_score – общая оценка университета, от 0 до 100, 1402 отсутствующих значения
- num\_students – число студентов, 59 отсутствующих значений
- student\_staff\_ratio – число сотрудников\*100/число студентов, 59 отсутствующих значений
- international\_students – процент заграничных студентов, 67 отсутствующих значений
- female\_male\_ratio – запись вида “a:b” такая, что  $a + b = 100$ , 233 отсутствующих значений
- year – дата сбора данных

# Data Analysis

Категориальные переменные, номинальные: university\_name, country

Количественные переменные, дискретные: teaching, international, research, citations, income, total\_score, international\_students, student\_staff\_ratio, female\_male\_ratio

Количественные переменные, непрерывные: num\_students, year

# Data Analysis

## После преобразований

#	Column	Non-Null Count	Dtype
0	world_rank	2601 non-null	int64
1	university_name	2601 non-null	object
2	country	2601 non-null	object
3	teaching	2601 non-null	float64
4	international	2592 non-null	float64
5	research	2601 non-null	float64
6	citations	2601 non-null	float64
7	income	2383 non-null	float64
8	total_score	1201 non-null	float64
9	num_students	2542 non-null	float64
10	student_staff_ratio	2542 non-null	float64
11	international_students	2534 non-null	float64
12	female_male_ratio	2365 non-null	float64
13	year	2601 non-null	int64

dtypes: float64(10), int64(2), object(2)

# Data Analysis

Создадим новую категориальную переменную, разбив интервальную переменную `world_rank` на группы `top`, `average` and `worst`

Мы выясним ключевые отличия между университетами, в зависимости от их расположения в рейтинге. Лучшими мы считаем университеты, которые имеют номер в мире до  $0.2 * \text{максимальный-номер}$ ; средними - от  $0.2 * \text{максимальный-номер}$  до  $0.8 * \text{максимальный-номер}$ ; худшими - от  $0.8 * \text{максимальный-номер}$



# Data Analysis

- **Задачи:**

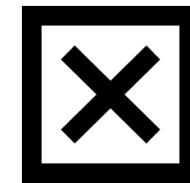
1. Выявить схожести и различия между университетами, расположенными в разных частях рейтинга.
2. Выявить связь рейтинга с локацией университетов.

- **Гипотезы, которые мы проверим:**

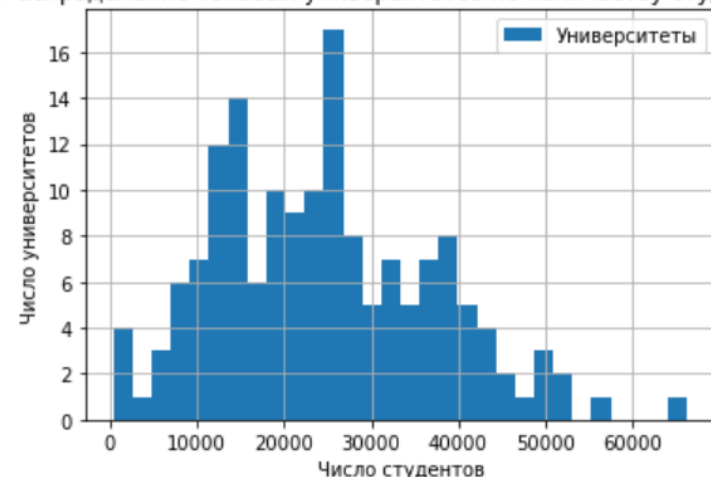
1. В топовых университетах в среднем меньше студентов, чем в остальных.
2. Самый большой доход у топовых университетов.
3. Соотношение количества женщин и мужчин не связано с категорией университета.
4. Больше всего крутых университетов в Англии или Америке, то есть, у них наибольший `total_score`.
5. Наибольшее число интернациональных студентов в Германии (по причине бесплатного образования)



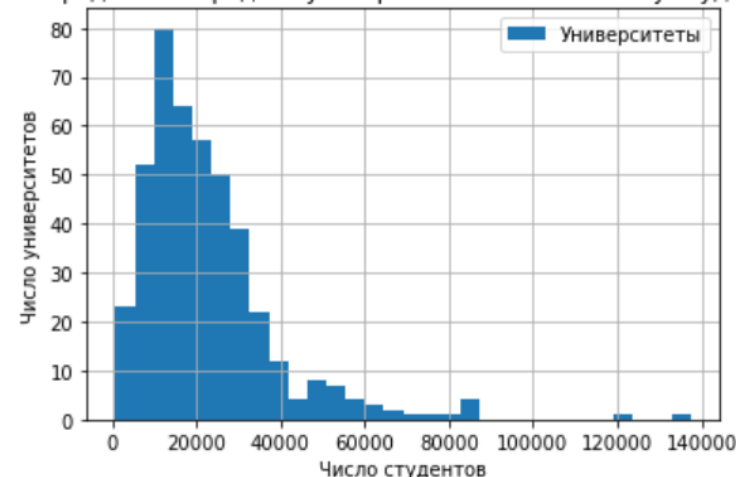
Гипотеза: в топовых университетах в среднем меньше студентов, чем в остальных.



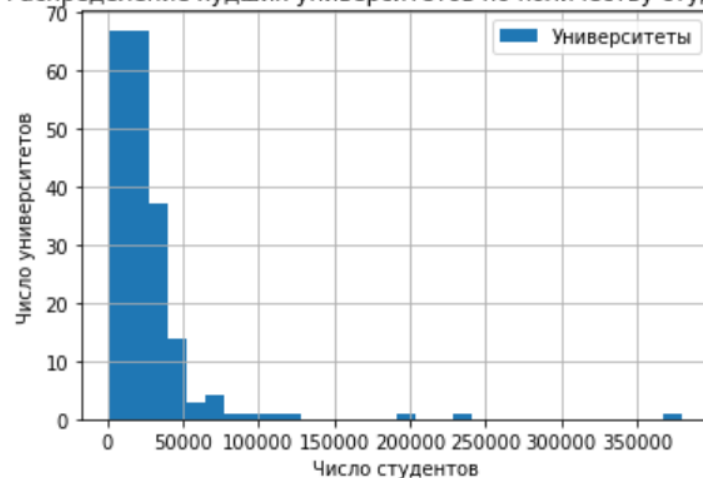
Распределение топовых университетов по количеству студентов



Распределение средних университетов по количеству студентов



Распределение худших университетов по количеству студентов



```
university_rank
average    22371.616972
top        24649.810127
worst      27694.412060
```

Гипотеза не подтвердилась.



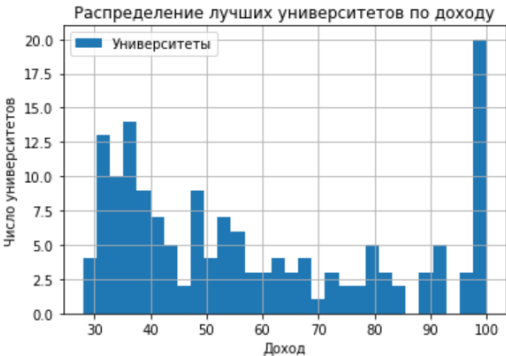
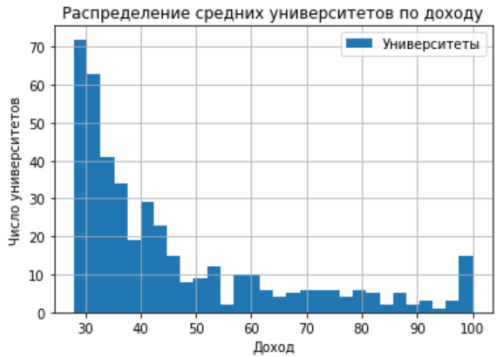
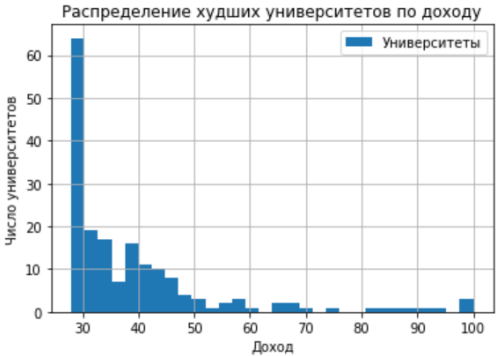
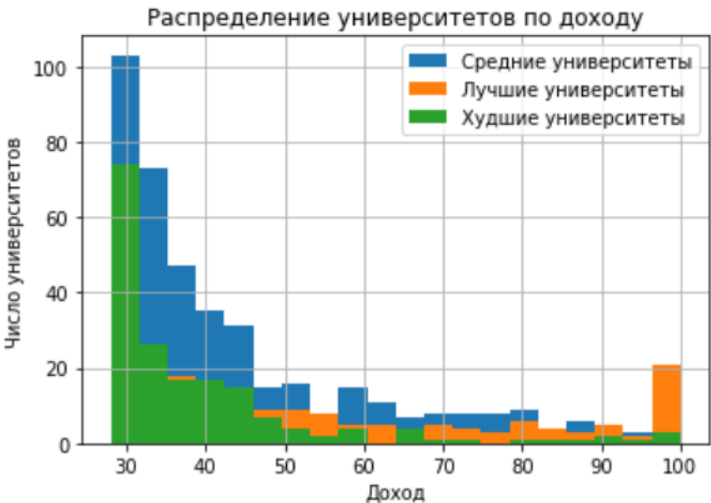
# Гипотеза: самый большой доход у топовых университетов.



	1813	1822	1831	1837	1844
world_rank	11	20	29	35	42
university_rank	top	top	top	top	top
university_name	Johns Hopkins University	Duke University	LMU Munich	KU Leuven	Peking University
income	100	100	100	100	100

...

	2512	2523	2531	2536	2571
	800	800	800	800	800
	worst	worst	worst	worst	worst
	University of Nairobi	Northwestern Polytechnical University	University of Paris North – Paris 13	Pontifical Catholic University of Paraná	Taras Shevchenko National University of Kyiv
	28	28	28	28	28



Гипотеза подтвердилась.

Гипотеза: соотношение женщин и мужчин не связано с категорией университета.



female_male_ratio	
university_rank	
average	111.594178
top	105.903084
worst	102.620313

	count	mean	std	min	25%	50%	75%	max
university_rank								
average	409.0	111.594178	47.242806	13.636364	85.185185	112.765957	138.095238	354.545455
top	144.0	105.903084	34.435731	21.951220	92.307692	108.333333	123.484848	233.333333
worst	182.0	102.620313	54.819350	1.010101	61.951348	92.307692	127.272727	354.545455

```
df['world_rank'].corr(df['female_male_ratio'])  
  
0.007782859939788572
```

Гипотеза подтвердилась.

Гипотеза: больше всего университетов с большим total\_score в Англии или Америке. Для того чтобы это проверить посмотрим какие страны в сумме обладают наибольшим значением.

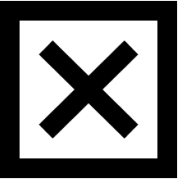


total_score	
country	
Australia	513.9
Austria	54.2
Belgium	232.1
Canada	368.7
China	70.0
Denmark	169.8
Finland	61.9
France	232.8
Germany	1049.9
Hong Kong	125.5
Israel	50.5
Italy	156.4
Luxembourg	49.4
Netherlands	733.7

New Zealand	51.0
Norway	50.1
Republic of Ireland	52.0
Russian Federation	51.9
Singapore	147.4
South Africa	56.1
South Korea	162.7
Spain	155.9
Sweden	245.9
Switzerland	391.6
Taiwan	51.1
United Kingdom	1972.0
United States of America	3769.6

Гипотеза  
подтвердилась.

# Гипотеза: Наибольшее число интернациональных студентов в Германии (по причине бесплатного образования).



international_students	
country	
United Kingdom	2076.0
United States of America	1778.0
Australia	791.0
Germany	471.0
France	466.0
...	...
Morocco	1.0
Kenya	1.0
Unisted States of America	0.0
Indonesia	0.0
Unted Kingdom	0.0

72 rows × 1 columns

```
new_df = pd.pivot_table(df,
                        values='international_students',
                        index=['country'],
                        aggfunc=np.sum)
new_df = new_df.sort_values(by=['international_students'], ascending=False)
new_df
```

Германия вошла в топ-5, но, видимо, не дотянула по причине меньшего количества университетов.

## Гипотеза не подтвердилась.

## Итоги:

- Университеты имеют разное число студентов, вне зависимости от их позиции в рейтинге.
- Доход университета имеет связь с его положением при ранжировании.
- Соотношение числа мужчин и женщин в университетах не имеет связи с его уровнем.
- Наиболее успешными в плане образования оказались Америка и Великобритания, у них же и оказалось больше всего заграничных студентов.

# Web scraping



Цель: собрать данные о  
жилых комплексах на  
территории Москвы

Источник: [www.cian.ru](http://www.cian.ru)



ЦИАН



ЦИАН

# Web scraping



**Квартиры в новостройках (ЖК) в Москве**

сданные новостройки 639    новостройки 2020 года 517    новостройки без отделки 443    до 6 млн рублей 298  
строящиеся новостройки 588    новостройки 2019 года 416    новостройки комфорт 362    новостройки эконом 281


**Акции и скидки**  
Каталог предложений на новостройки

**Скоро в продаже**  
Новостройки на стадии котлована

**Старт продаж**  
Квартиры по стартовой цене

**Лидеры продаж**  
Большой спрос на квартиры

1 011 предложений На карте




**ЖК «Aquatoria (Акватория)»** Проверено ЦИАН  
Сделано в 2022 году, монолитный

**Беломорская 18 мин. пешком**  
Москва, САО, Левобережный, Ленинградское ш., 69  
Жилой комплекс «Aquatoria (Акватория)» возводится в Левобережном районе Северного административного округа Москвы. Пять корпусов бизнес-класса расположены на берегу Химкинского водохранилища, их объединит...

2-комн. от 52,52 м² 23,5–25 млн ₽	3-комн. от 67,34 м² 29,4–53,65 млн ₽	4-комн. от 115,43 м² 46,35–89,85 млн ₽
5-комн. от 210,89 м² 96,5–138,5 млн ₽		

**Контакты застройщика** В избранное 41 квартира от застройщика · 23 от агентов



**ЖК «Триколор»**  
Сделано в 2022 году, кирпичный

**Ростокино 14 мин. пешком**  
Москва, СВАО, Ростокино, просп. Мира, 188  
Комплекс находится в районе Ростокино СВАО на пересечении Ростокинской улицы с Проспектом Мира, на территории бывшего Деревообрабатывающего комбината №17, вблизи парка Ростокинского акведука на берегу...

2-комн. от 97,20 м² 30,52–40,77 млн ₽	3-комн. от 139 м² 33,54–59,55 млн ₽
--	--

**Контакты застройщика** В избранное 94 квартиры от застройщика · 100 от агентов

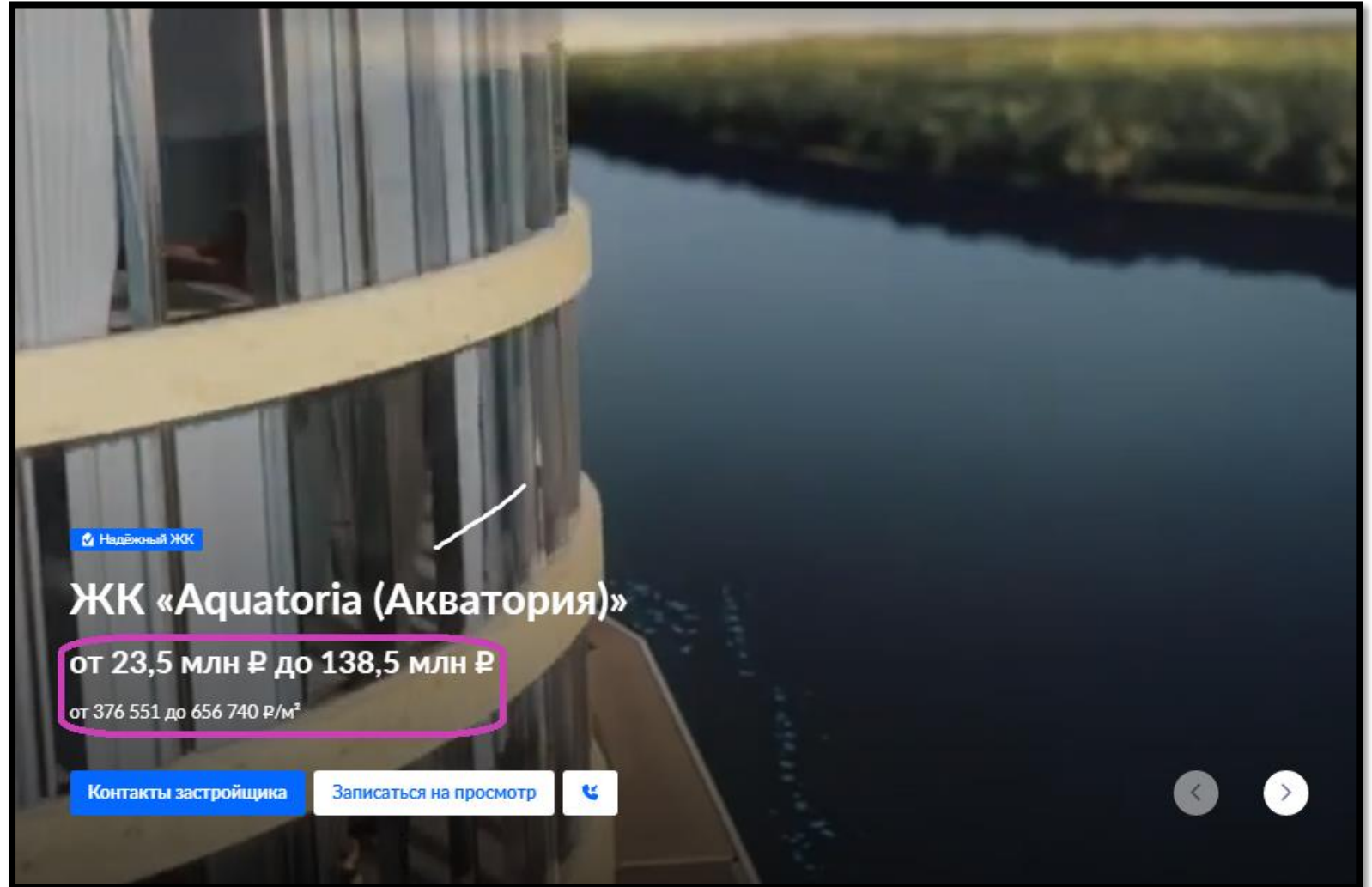
Начальная  
страница:  
[www.cian.ru/  
novostroyki-2020/](http://www.cian.ru/novostroyki-2020/)

# Web scraping



Целевые данные:

- Название
- Цена
- Ближайшее станция метро
- Информация о конструкции
- Количество квартир в продаже















# Web scraping



Целевые данные:

- Название
- Цена
- Ближайшее станция метро
- Информация о конструкции
- Количество квартир в продаже

 м. Беломорская  18 мин.

 Срок сдачи 2021–2023 	 Класс Бизнес	 Этажность от 11 до 21 
 Тип дома Монолитный	 Высота потолков 2,95 м	 Варианты отделки Чистовая

# Web scraping







Целевые данные:



- Название
- Цена
- Ближайшее станция метро
- Информация о конструкции
- Количество квартир в продаже






129 квартир от застройщика в ЖК «Дом Лаврушинский»

Все корпуса

Строение 1    
Сдача во II-кв. 2024

Строение 2    
Сдача во II-кв. 2024

Строение 3    
Сдача во II-кв. 2024

1-комнатные	от 65 м²	109,2–149,6 млн ₽	7 предложений 
2-комнатные	от 86 м²	129–582,2 млн ₽	41 предложение 
3-комнатные	от 125 м²	203,1–633,9 млн ₽	40 предложений 
4-комнатные	от 196 м²	357,7–1729,5 млн ₽	27 предложений 
5-комнатные	от 227 м²	531,8–1615,9 млн ₽	14 предложений 

Всего 129 квартир в ЖК

# Web scraping



Итоговый результат:  
Таблица на 190+ позиций

Name	ClosestStation	AvgPrice	FullPriceMin	FullPriceMax
ЖК «Aquatoria (Акватория)»	м. Беломорская	55,7	23,5	138,5
ЖК «Триколор»	м. Ростокино	32,7	30,52	59,55
ЖК «Композиция №24»	м. Шаболовская	94,9	91,98	99,99
ЖК «Символ»	м. Площадь Ильича	26	13,03	64,54
Название	Ближайшее метро	Средняя цена (млн. Руб.)	От	До
			Общая цена (млн. Руб.):	

Цена за метр (тыс. Руб.):		Кол-во квартир с разной планировкой					
От	До						
PricePerMeterMin	PricePerMeterMax	1room	2rooms	3rooms	4rooms	5rooms	studios
376	656	0	6	8	21	4	0
200	372	0	49	45	0	0	0
619	627	0	0	0	3	0	0
290	487	27	105	37	14	2	4

Class	DateOfConstruction	FloorsNumber	Type
Бизнес	2015	10-58	Монолитно-кирпичный
Бизнес	2021	8	Монолитный
Премиум	2023	10-11	Монолитный
Класс	Дата постройки	Кол-во этажей	Тип дома