|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме\_\_\_\_\_\_\_\_\_\_\_\_\_ «Рейтинг лучших университетов 2017-2022 гг.» \_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5-53Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** М.О. Пономарева

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** К.Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2022 г.*

**Аннотация**

В задании необходимо проанализировать базу данных “ QS World University Rankings” и очистить от не нужных данных, также нужно выдвинуть гипотезы, проанализировать их и составить графики по данным гипотезам. В ходе работы будут использоваться такие библиотеки как: «Pandas» и «Seaborn». Анализ будет проводиться через программу «Jupyter Python».

Также будут проводиться корреляционный анализ и удаление дубликатов, очистка данных. Будет производиться анализ гипотез по известным данным и построены соответствующие графики и схемы.

В работе будут рассмотрены зависимости места университета в рейтинге от множества факторов. И в конце концов, будет сделан вывод о проделанной работе с данным датасетом.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 3](#_Toc122024818)

[1. Описание данных 5](#_Toc122024819)

[2. Анализ данных 7](#_Toc122024820)

[3. Исследовательский анализ данных 12](#_Toc122024821)

[3.1. Вычисление среднего рейтинга у университетов разных типов 12](#_Toc122024822)

[3.2. Вычисление отношения показателя соотношения количества студентов на одного преподавателя к месту в рейтинге 13](#_Toc122024823)

[3.3. Вычисление количества университетов, попавших в рейтинг в каждой стране 13](#_Toc122024824)

[3.4. Анализ корреляции 14](#_Toc122024825)

[4. Общий вывод 16](#_Toc122024826)

[ЗАКЛЮЧЕНИЕ 17](#_Toc122024827)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 18](#_Toc122024828)

# **ВВЕДЕНИЕ**

Необходимо исследовать базу данных и выявить закономерности места в рейтинге университетов.

Цели:

1. определение данных;
2. формулирование гипотез;
3. загрузка данных в Python;
4. проверка данных;
5. очистка данных;
6. преобразование данных;
7. выбор данных для анализа;
8. агрегирование данных;
9. визуализация данных;
10. подтверждение или опровержение поставленных гипотез;
11. формулирование ограничений и выводов.

**Определение данных для анализа**

В качестве данных для анализа был выбран Dataset «QS World University Rankings».

В датасете представлен рейтинг мировых университетов с 2017 по 2022 гг. с информацией о каждом университете и подробным описанием его характеристик.

# **Описание данных**

Для анализа были собраны данные о всех страховках, включенных в датасет. В наборе данных содержатся:

1. university – название университета;
2. year – год попадания в рейтинг;
3. rank\_display – место в рейтинге;
4. score – количество баллов полученное университетом;
5. link – ссылка на университет на сайте QS;
6. country – страна, в которой университет располагается;
7. city – город, в которой университет располагается;
8. region – континент, на котором университет располагается;
9. logo – ссылка на логотип университета;
10. type – тип университета (частный или публичный);
11. research\_output – качество, с которым проводилось исследование;
12. student\_faculty\_ratio – соотношение студентов и преподавателей;
13. international\_students – число иностранных студентов в университете;
14. size – размер университета;
15. faculty\_count – количество профессорско-преподавательского состава в университете.

**Формулирование гипотез**

Гипотеза 1:

Частные университеты имеют более высокий рейтинг, нежели публичные.

Гипотеза 2:

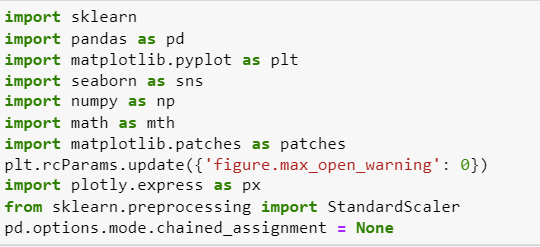
Чем ниже показатель соотношения студентов и преподавателей, тем более высокую строчку рейтинга занимает университет.

Гипотеза 3:

США имеет наибольшее количество университетов, попавших в рейтинг.

# **Анализ данных**

Загружаем датасет и подключаем необходимые библиотеки:



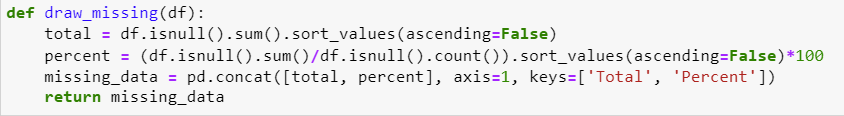
**Загрузка данных в Python**

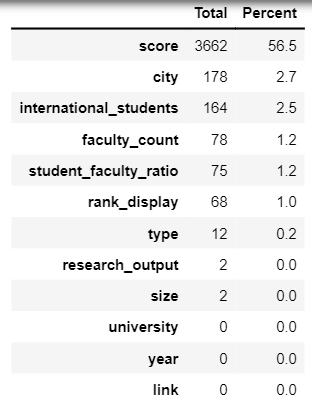
Получим информацию о датасете:



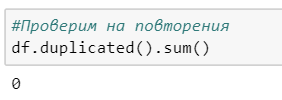
**Проверка данных**

Проверим набор на наличие пропусков, для этого воспользуемся функцией draw\_missing:





Проверим набор на наличие дубликатов, для этого воспользуемся функцией duplicated:



Таким образом, имеем некоторое наличие пропусков и полное отсутствие дубликатов.

**Очистка данных**

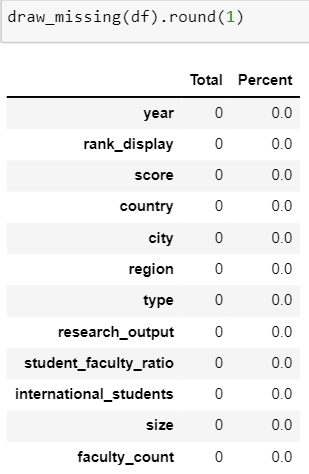
Во всех столбцах количество пропусков не превышает 20% (за исключением столбца score), поэтому мы можем позволить себе их удалить, используя функцию dropna [7]:



Мы производим анализ по лучшим университетам исходя из их места в рейтинге и количества баллов, однако университеты, находящиеся на последних местах, не имеют четкого места и имеют пустые значения на месте баллов, поэтому не пригодны для анализа. Удалим их [6]:



Снова проверим количество пропусков.

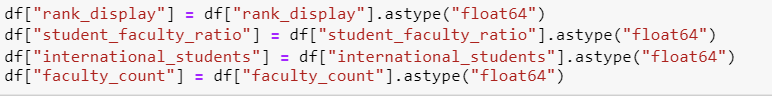


**Вывод:**

Т.к. столбец, в котором были пропущены значения являлся отправным для анализа заполнение являлось невозможным, поэтому было принято решение его удалить.

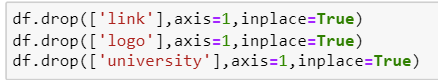
**Преобразование данных**

Там, где это возможно преобразуем тип string в тип float:



**Выбор данных для анализа**

Заметим, что данные содержащиеся в link, logo и university - уникальные и бесполезны для анализа, поэтому удалим их из датафрейма:

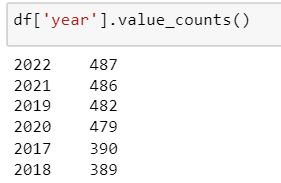


**Результат подготовки данных для анализа**

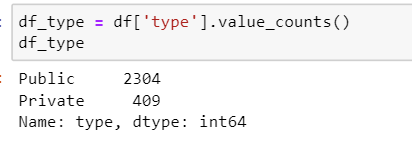
****

**Агрегирование данных**

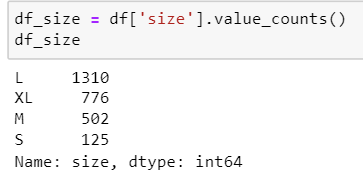
Посчитаем количество строк (мест в рейтинге) за каждый год, используя функцию value\_counts ():



Посчитаем количество частных и публичных университетов:

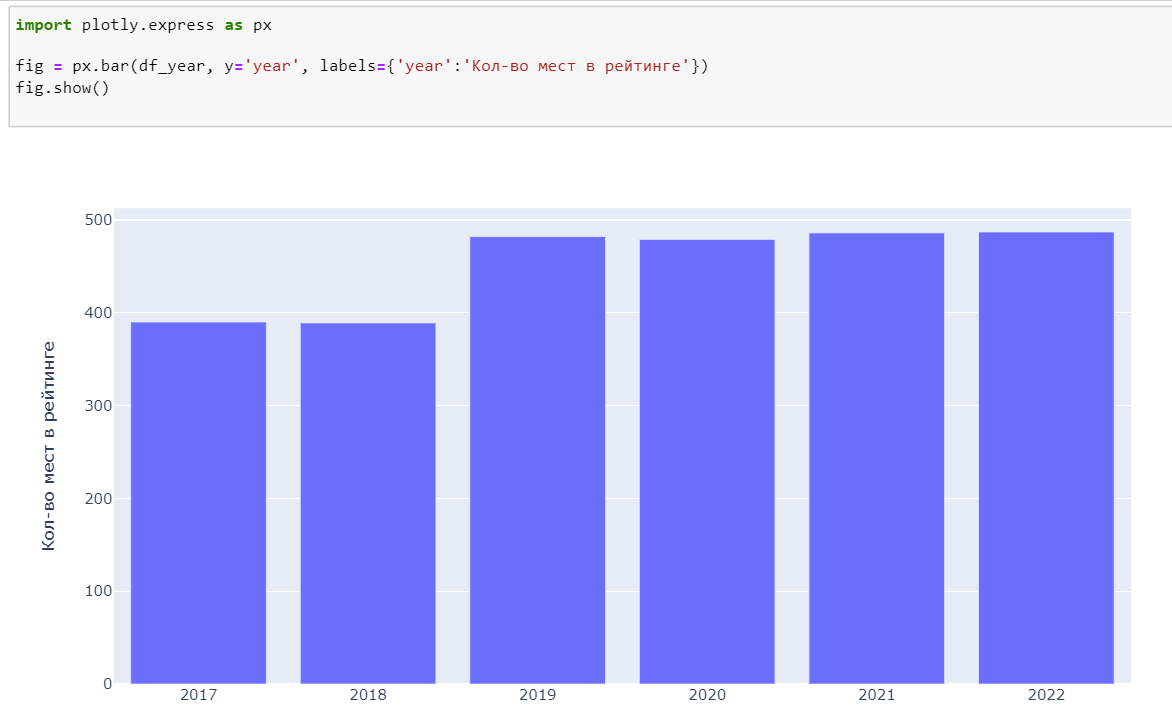


Посчитаем количество университетов разного размера:

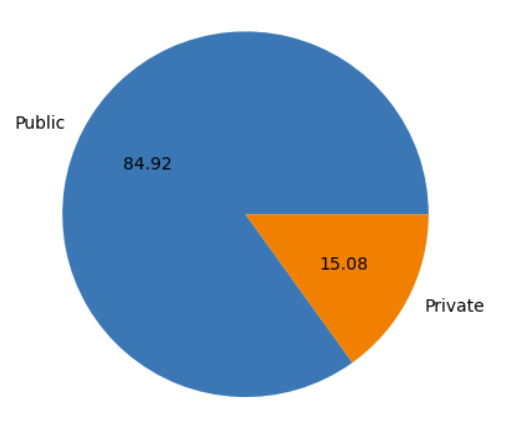


**Визуализация данных**

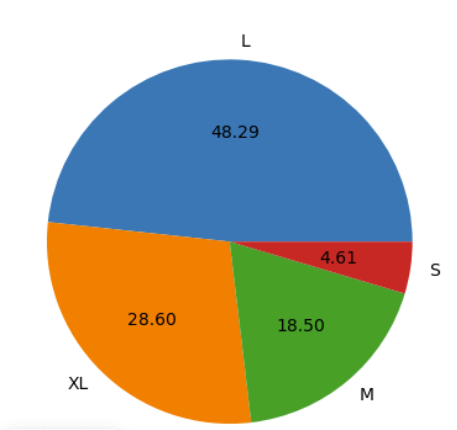
Визуализация количества мест в рейтинге за каждый год:



Визуализация соотношения частных и публичных университетов:



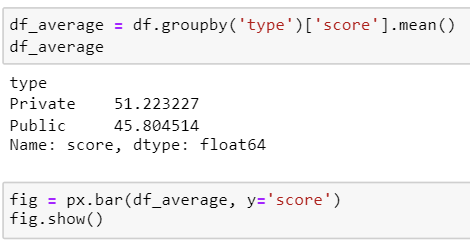
Визуализация количества университетов разного размера:



# **Исследовательский анализ данных**

* 1. **Вычисление среднего рейтинга у университетов разных типов**

Известно, что университеты делятся на частные и публичные (государственные). Давайте выясним какие из них имеют в среднем более высокий рейтинг.



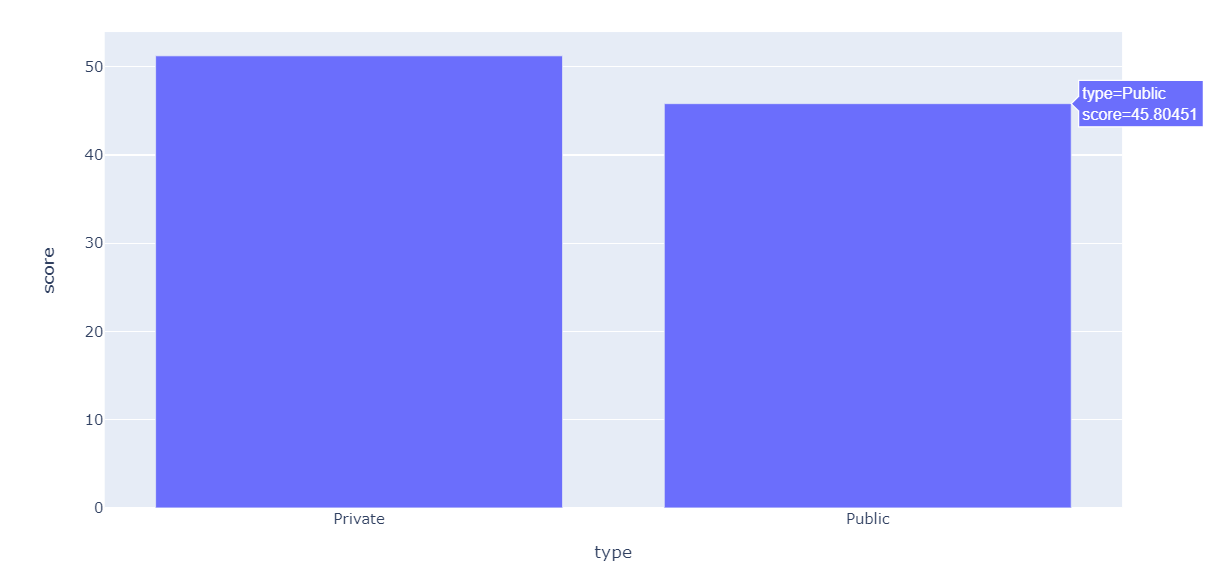


Рисунок 3.1 – Средний рейтинг университетов разных типов

**Вывод**:

По графику на рисунке 3.1 можно заметить, что пусть с небольшим отрывом, но частные университеты имеют более высокий рейтинг нежели публичные.

Гипотеза 1 – подтверждена.

* 1. **Вычисление отношения показателя соотношения количества студентов на одного преподавателя к месту в рейтинге**

В каждом университете имеется разный показатель соотношения количества студентов на одного преподавателя. Это зависит от направления, финансирования и т.д. Выясним как этот показатель влияет на рейтинг:



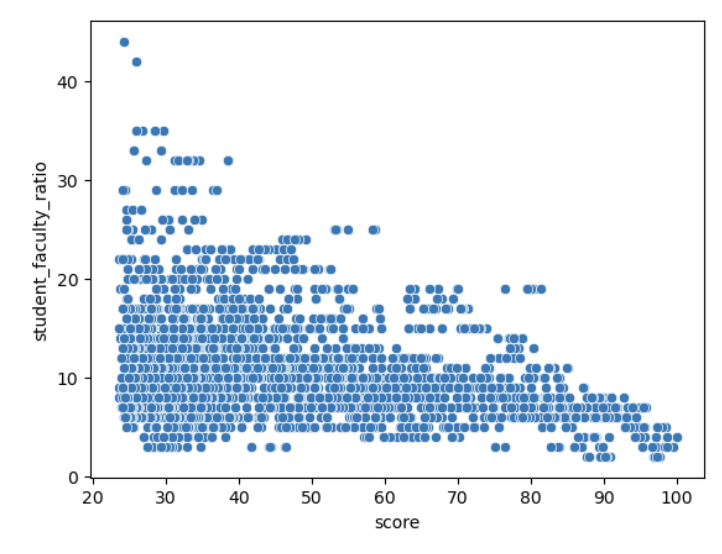


Рисунок 3.2 – Отношение показателя соотношения количества студентов на одного преподавателя к месту в рейтинге

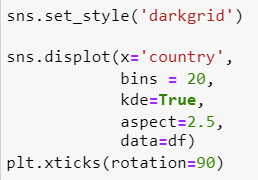
**Вывод:**

По графику на рисунке 23 можно заметить, что чем ниже показатель соотношения количества студентов на одного преподавателя, тем выше место университета находится в рейтинге.

Гипотеза 2 – подтверждена.

* 1. **Вычисление количества университетов, попавших в рейтинг в каждой стране**

В каждой стране разное количество высших учебных заведений, это зависит от ее размеров, плотности населения и отношения в стране к высшему образованию. Выясним в каких странах наибольшее количество университетов, попавших в рейтинг:



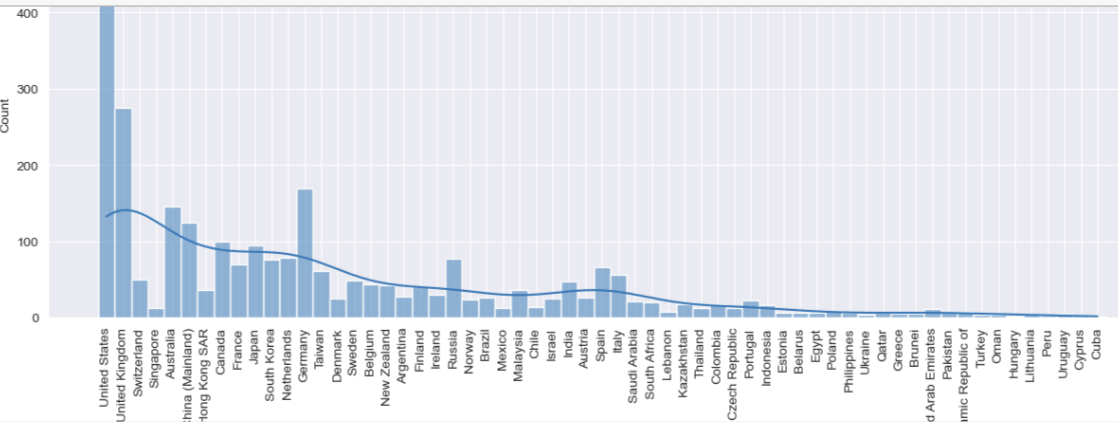


Рисунок 3.3 – Количество университетов попавших в рейтинг в каждой стране

Вывод:

По графику на рисунке 3.3 явно видно, что страной с наибольшим количеством университетов, попавших в рейтинг, является США.

Гипотеза 3 – подтверждена.

* 1. **Анализ корреляции**

Создадим матрицу корреляций [5]:

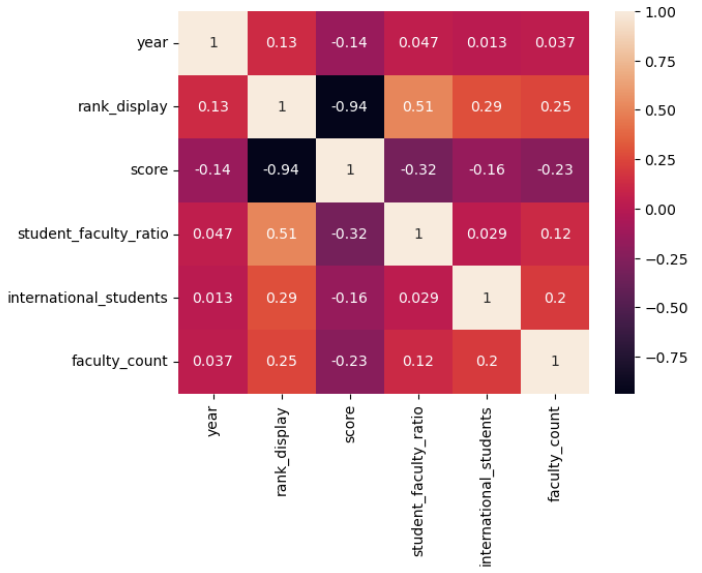
****

Рисунок 3.4 – Диаграмма корреляции

Коэффициент корреляции Пирсона измеряет линейную связь между переменными. Его значение можно интерпретировать так:

r>0,01≤0,29 – слабая положительная связь,

r>0,30≤0,69 – умеренная положительная связь,

r>0,70≤1,00 – сильная положительная связь,

r>-0,01≤-0,29 – слабая отрицательная связь,

r>-0,30≤-0,69 – умеренная отрицательная связь,

r>-0,70≤-1,00 – сильная отрицательная связь.

Проанализируем полученные результаты. Коэффициент корреляции Пирсона равный единице стоит на пересечении переменных с самим собой. Остальные коэффициенты корреляции имеют либо слабую отрицательную связь, либо слабо положительную, за исключением значения между score и rank\_display.

Т.к. score является более точным значением, удалим столбец rank\_display.

**Вывод корреляционного анализа:**

Подводя итог корреляционного анализа, можно сделать вывод, что после удаления столбца score избыточности нет.

# **Общий вывод**

Перед анализом была произведена очистка данных и предварительная визуализация данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания в предметной области «QS World University Rankings».

В ходе анализа было выявлено:

1. Частные университеты имеют более высокий рейтинг, нежели публичные.
2. Чем ниже показатель соотношения студентов и преподавателей, тем более высокую строчку рейтинга занимает университет.
3. США имеет наибольшее количество университетов, попавших в рейтинг.

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ базы данных «QS World University Rankings», выявлены важные статистические данные, которые могут улучшить качество работы университетов.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas» и «Seaborn» на базе языка программирования Python [1], [2], [3]. Также мы пользовались «Jupyter Python» в нём мы писали код и анализировали диаграммы [4]. В ходе работы я получил новые навыки в анализе данных и познакомился с вышеперечисленными программами.

В ходе анализа выявлено, что рейтинг университета во многом зависит от показателя соотношения количества студентов на одного преподавателя, а также от того частным или публичным является университет.

# 

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Методические указания по программному обеспечению «Pandas»;
2. Методические указания по программному обеспечению «Seaborn»;
3. «Pandas. Работа с данными» (2020), Автор: Абдрахманов М. И.;
4. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi»;
5. <https://devpractice.ru>;
6. <https://coderoad.wiki>;
7. <https://www.delftstack.com>.