|  |  |
| --- | --- |
|  | **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 «Best Universities in the United Kingdom»\_\_\_\_\_\_\_\_\_\_\_\_

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

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

Студент ИУ5И-51Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** А. Кашима

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

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

(Подпись, дата) (И.О.Фамилия)

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

(Подпись, дата) (И.О.Фамилия)

*2024 г.*

**Аннотация**

В задани необходимо проанализировать базу данных “ Best Universities in the United Kingdom ” и очистить от не нужных данных, также 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, агрегирование данных, оптимизация память, проверка на наличие дубликатов и в случае положительного результата – очистка от лишних данных. Будет производиться анализ гипотез по известным данным и построены соответствующие графики и схемы.

В работе будут рассмотрены и проанализированы университеты Соединенного Королевства, обучение в них, удовлетворенность учащихся и другие факторы при выборе Университета для возможного поступления. И в конце концов, будет сделан вывод о проделанной работе с данным датасетом.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ4](#_Toc122082171)

[1. Описание данных5](#_Toc122082172)

[2. Изучение общей информации7](#_Toc122082173)

[3. Подготовка данных для работы10](#_Toc122082174)

[4. Исследовательский анализ данных15](#_Toc122082175)

[4.1 Распределение университетов по регионам15](#_Toc122082176)

[4.2 Распределение университетов по местоположению кампуса17](#_Toc122082177)

[Определение самых лучших университетов в миреError: Reference source not found](#_Toc122082178)

[5. Общий вывод21](#_Toc122082179)

[ЗАКЛЮЧЕНИЕ22](#_Toc122082180)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ23](#_Toc122082181)

# **ВВЕДЕНИЕ**

Необходимо исследовать базу данных и выявить закономерности рейтингов 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 «Best Universities in the United Kingdom».

В датасете представлен рейтинг университетов Соединенного Королевства с информацией о каждом университете.

# **Описание данных**

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

1. Название университета;
2. регион;
3. год основания;
4. девиз;
5. национальный ранг;
6. мировой рейтинг;
7. оценка мировых рейтингов;
8. минимальные баллы для поступления;
9. распределение иностранных студентов;
10. оценка студентов;
11. кол-во поступивших студентов (тыс.);
12. кол-во преподавательского состава;
13. тип управления университетом;
14. местоположение кампуса;
15. стоимость жизни в год;
16. широта;
17. долгота.

**Формулирование гипотез**

Гипотеза 1:

Больше всего университетов в Лондоне.

Гипотеза 2:

Больше всего университетов находится в пределах городской черты.

Гипотеза 3:

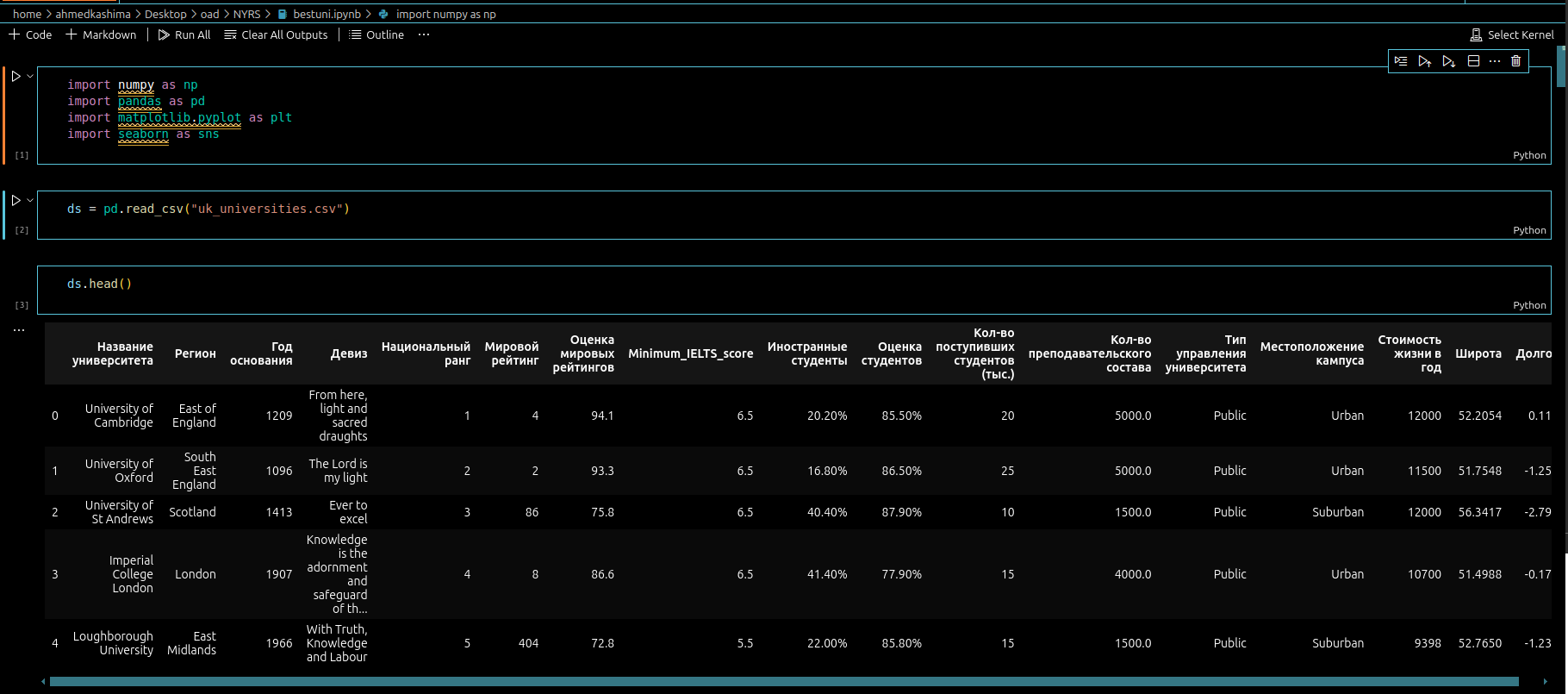
Самый лучший университет Соединенного Королевства – Оксфордский.

Гипотеза 4:

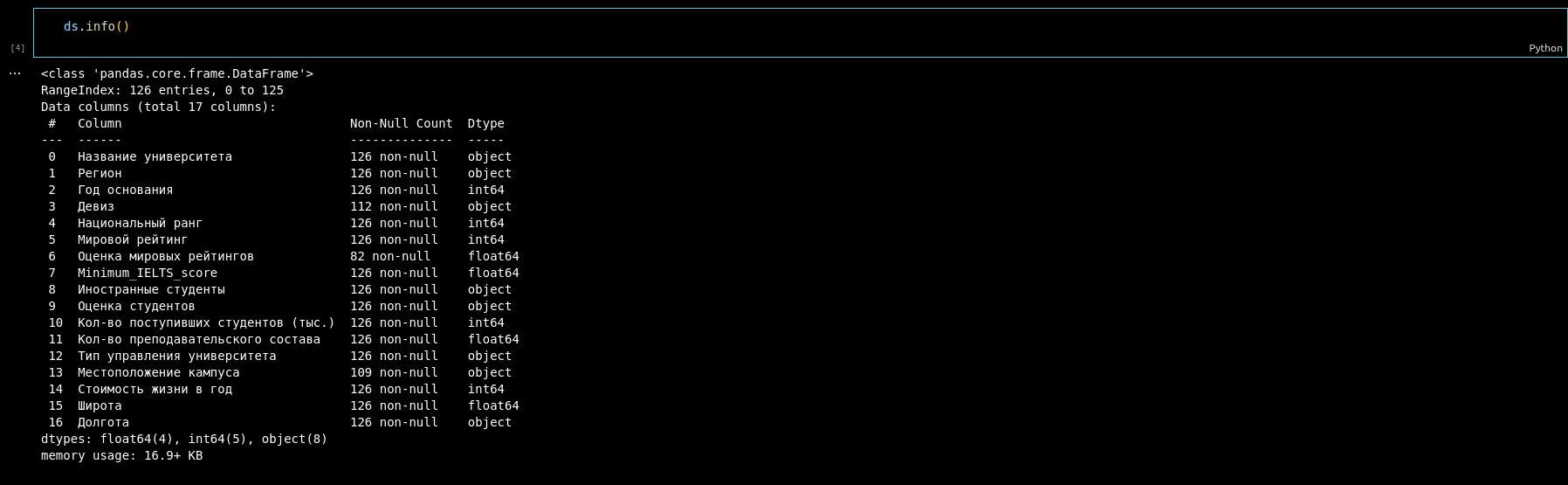
Самая большая стоимость жизни в Кэмбридже.

# **Изучение общей информации**

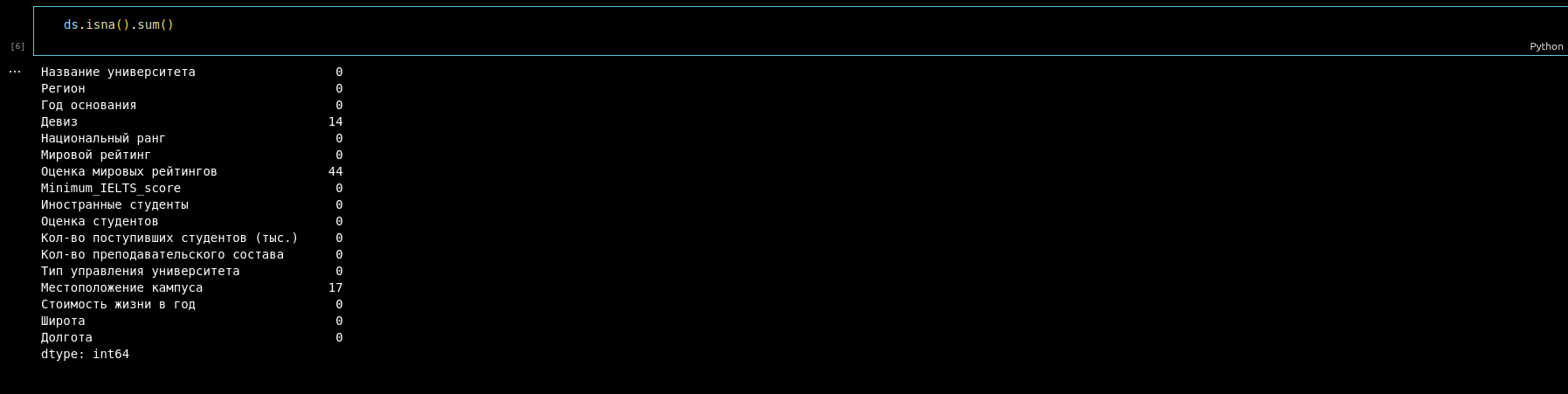
Загружаем датасет и подключаем необходимые библиотеки:



Получим информацию о датасете:



Посчитаем количество пустых значений:



Так как мы нашли строки с пустыми значениями, то удалим их из датасета с помощью команды:



Посчитаем снова количество пустых значений и посмотрим сколько всего строк осталось в датасете:

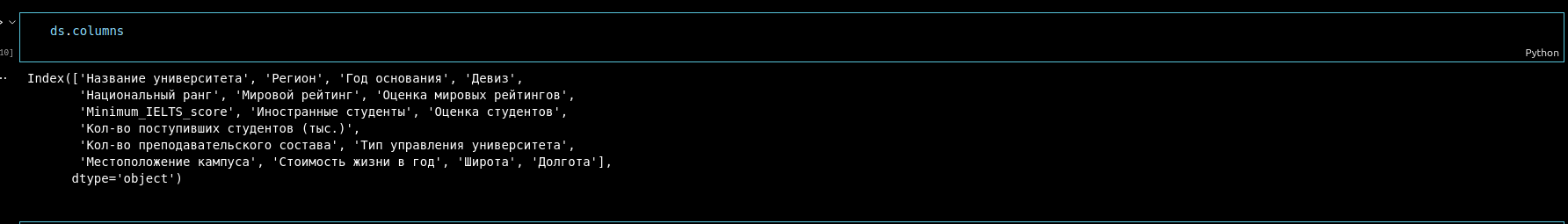


**Вывод:**

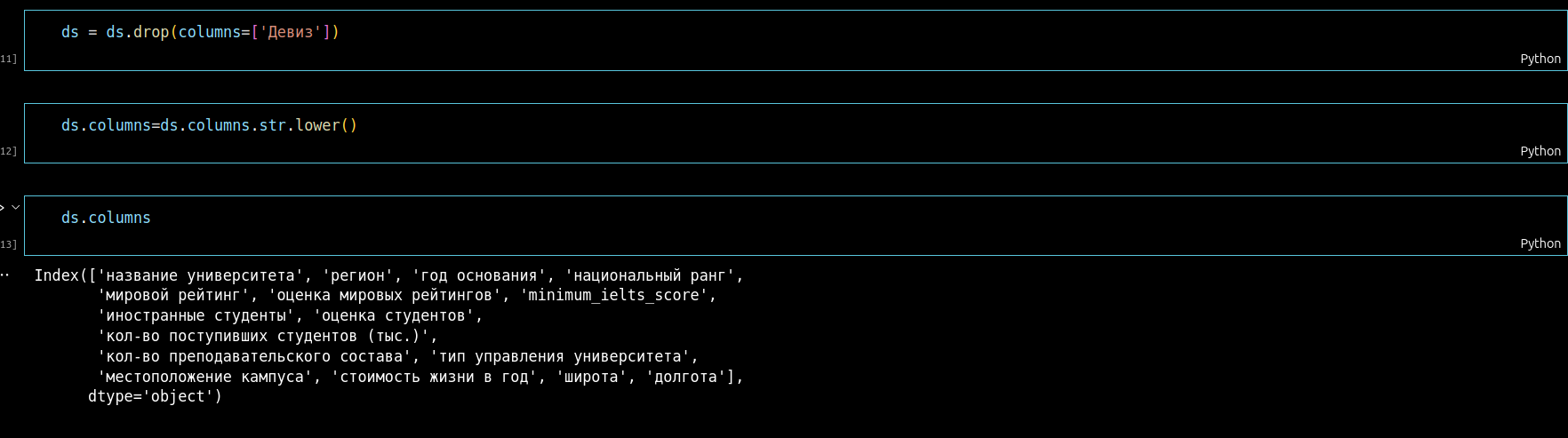
В начале работы с датасетом были пустые строчки, после определённых команд мы их удалили и в конечном итоге осталось всего 985861 строка в датасете. Пустые строки были удалены, так как нас интересуют конкретные значения в датасете, невозможно было заполнить пустые строки информации, потому что это данные действия могли бы привести к неверным или искаженным результатам.

# **Подготовка данных для работы**

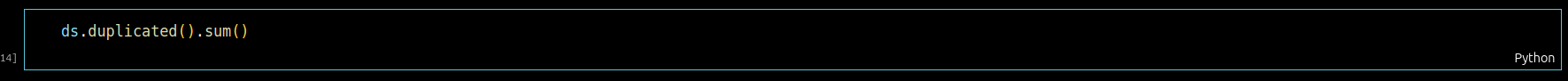
Проверим названия столбцов:



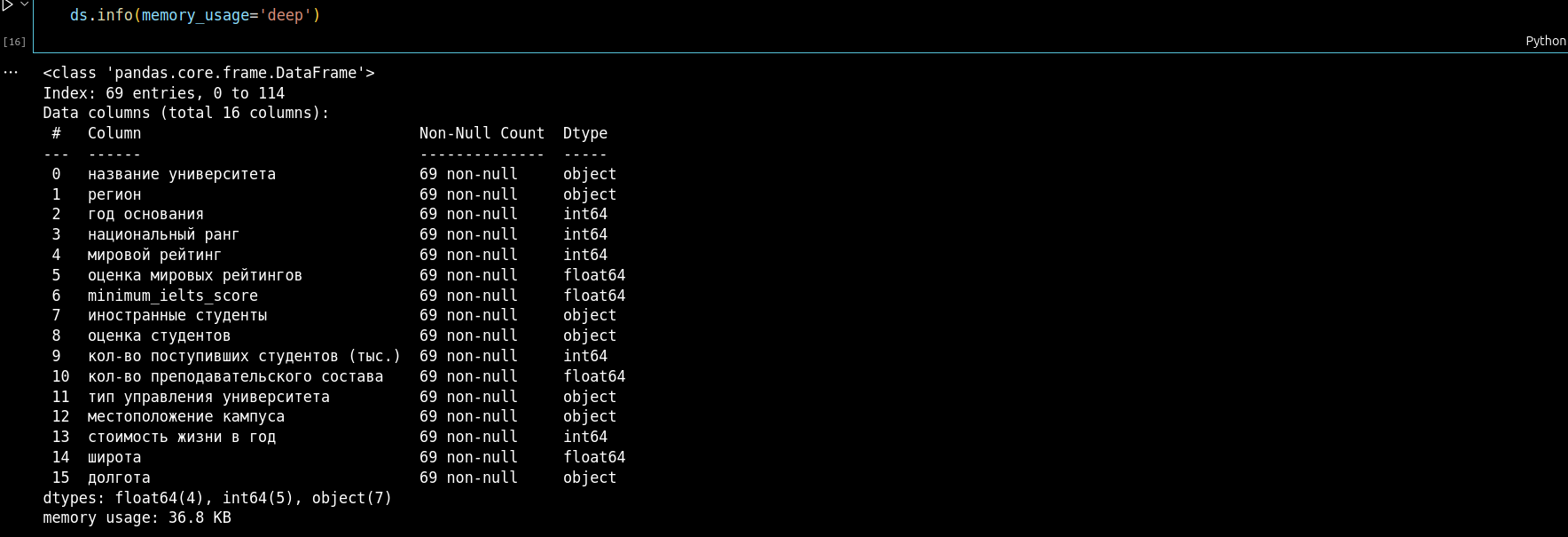
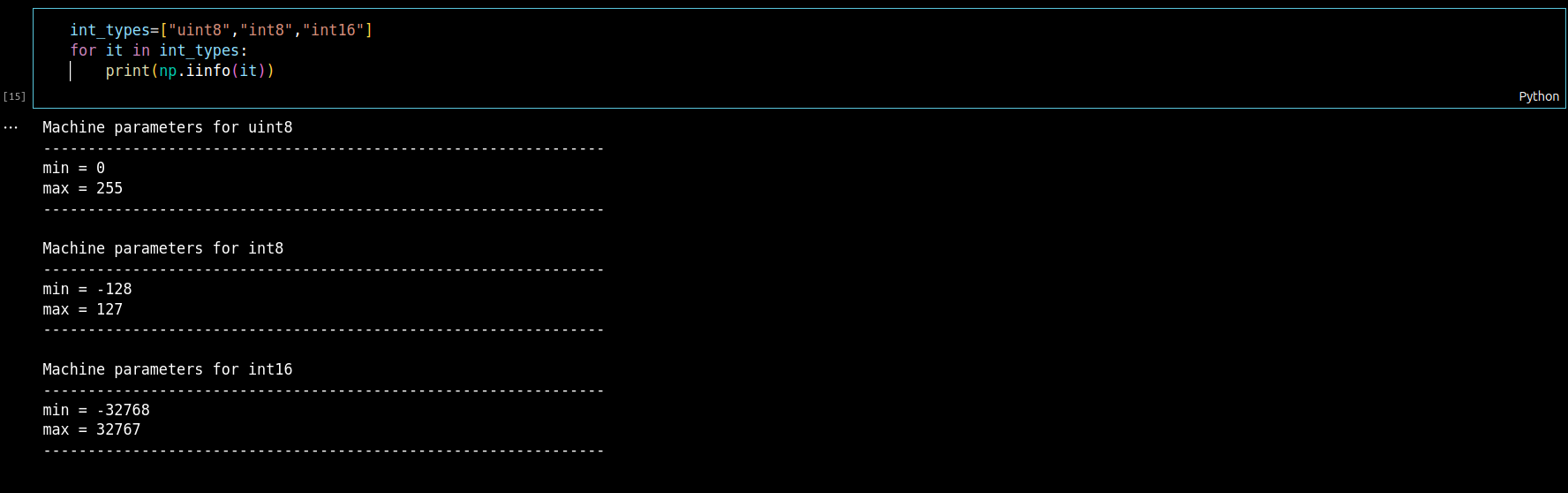
При работе с датасетом нам следует удалить колонку девиз, так как она не несёт в себе полезной для анализа информации. Удалим его:



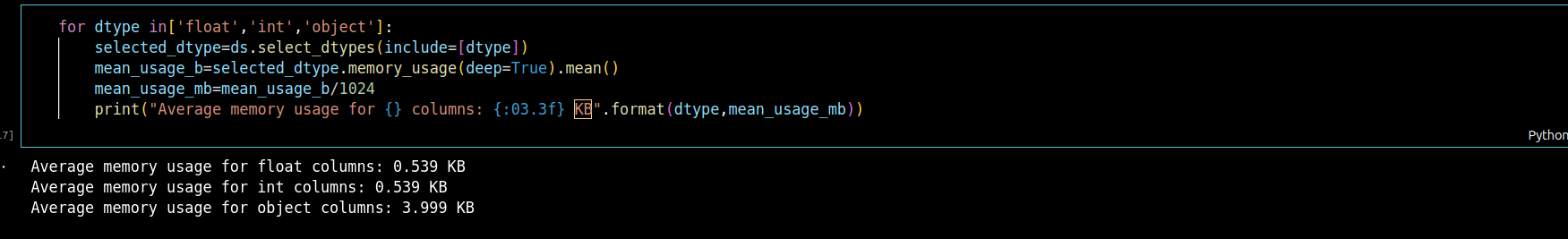
Проверим датасет на наличие дубликатов:



Оптимизируем память и изменим типы данных на более корректные:

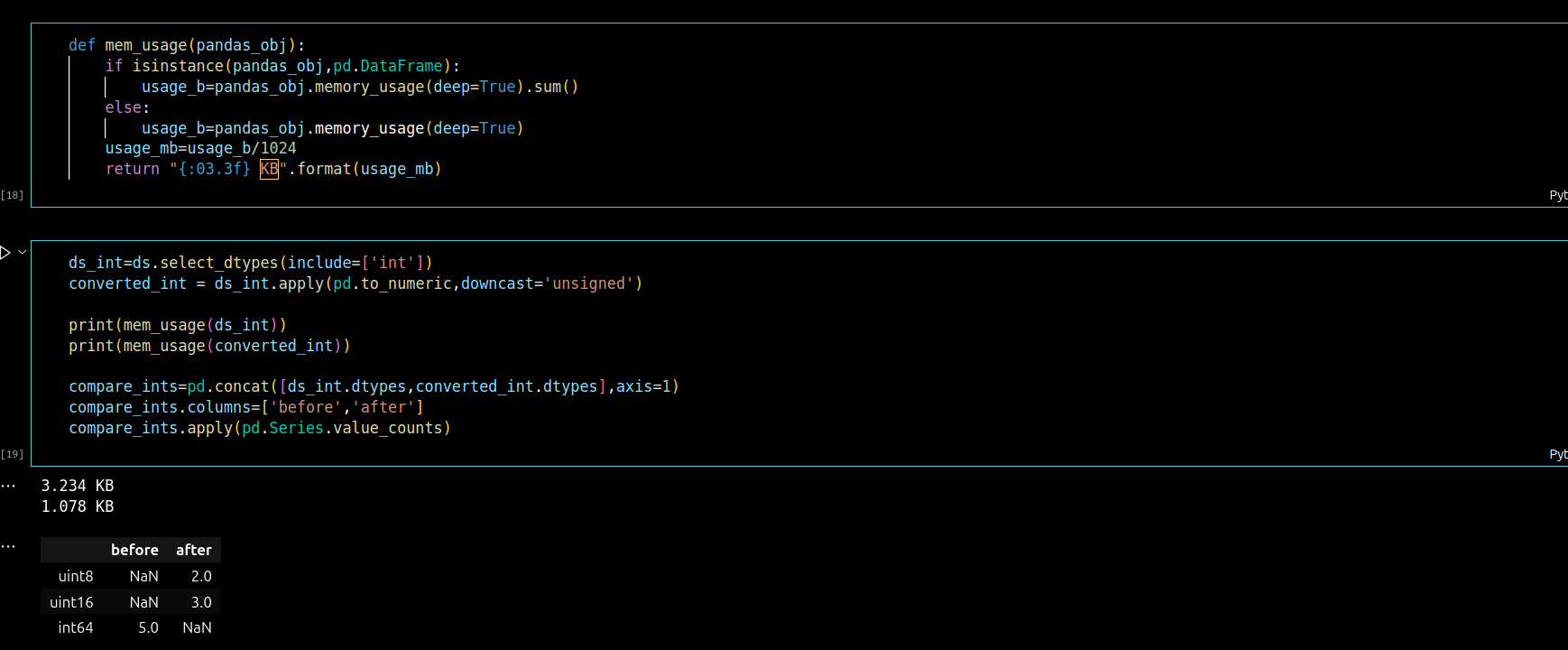


Так как данные разных типов хранятся раздельно, мы исследуем использование памяти разными типами данных. Начнём со среднего показателя использования памяти по разным типам данных:



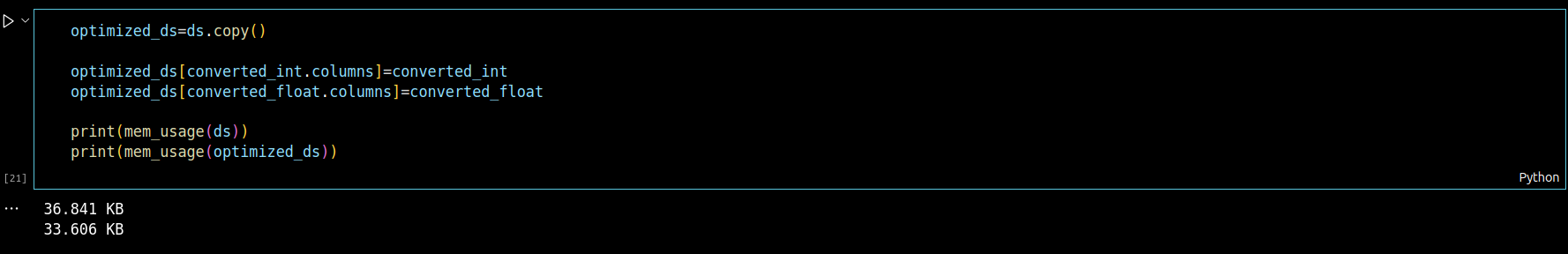
Эти сведения дают нам понять то, что большая часть памяти уходит на столбцы, хранящих объектные значения.

Создадим функцию, которая поможет нам сэкономить немного времени:



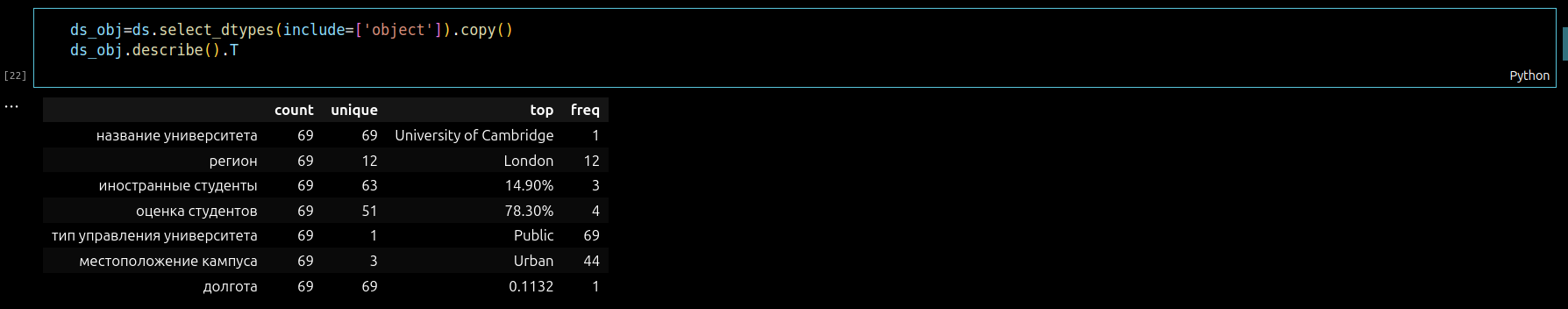
В результате можно видеть падение использования памяти int-данными с 3.234 KB до 1.078 KB, float-данными 2.695 KB до 1.617 KB, то есть — мы снизили потребление памяти.

Создадим копию исходного объекта Data Frame, используем эти оптимизированные числовые столбцы вместо тех, что присутствовали в нём изначально, и посмотрим на общий показатель использования памяти после оптимизации:

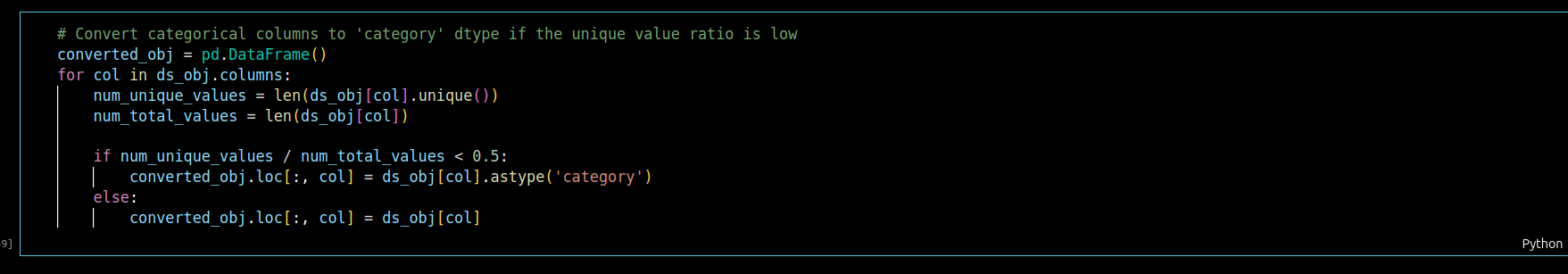


После уменьшения потребления памяти столбцами, хранящими числовые данные, нам удалось уменьшить потребление памяти на 2%. Для достижения еще большего результата оптимизируем хранение объектных типов.

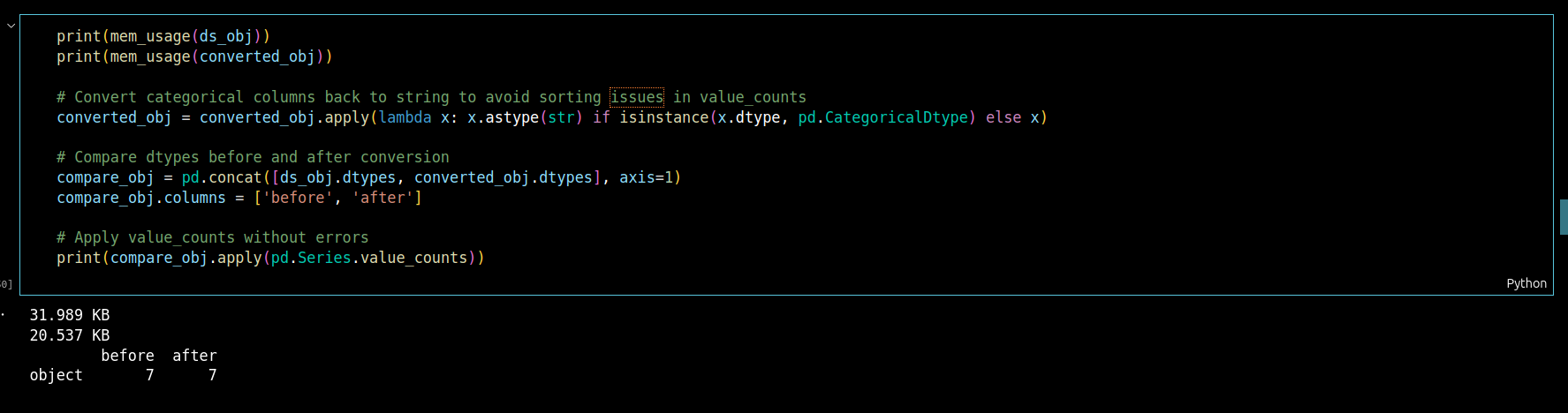
Для того чтобы понять, где именно мы сможем воспользоваться категориальными данными для снижения потребления памяти, выясним количество уникальных значений в столбцах, хранящих значения объектных типов:

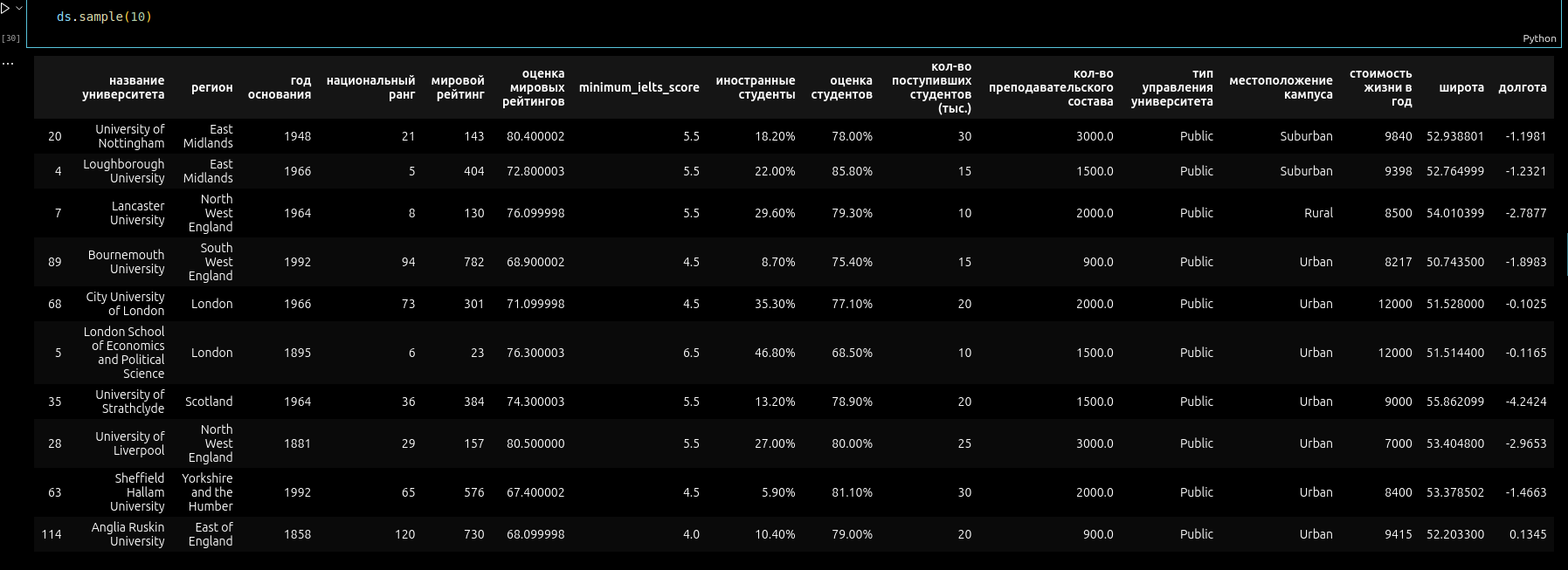
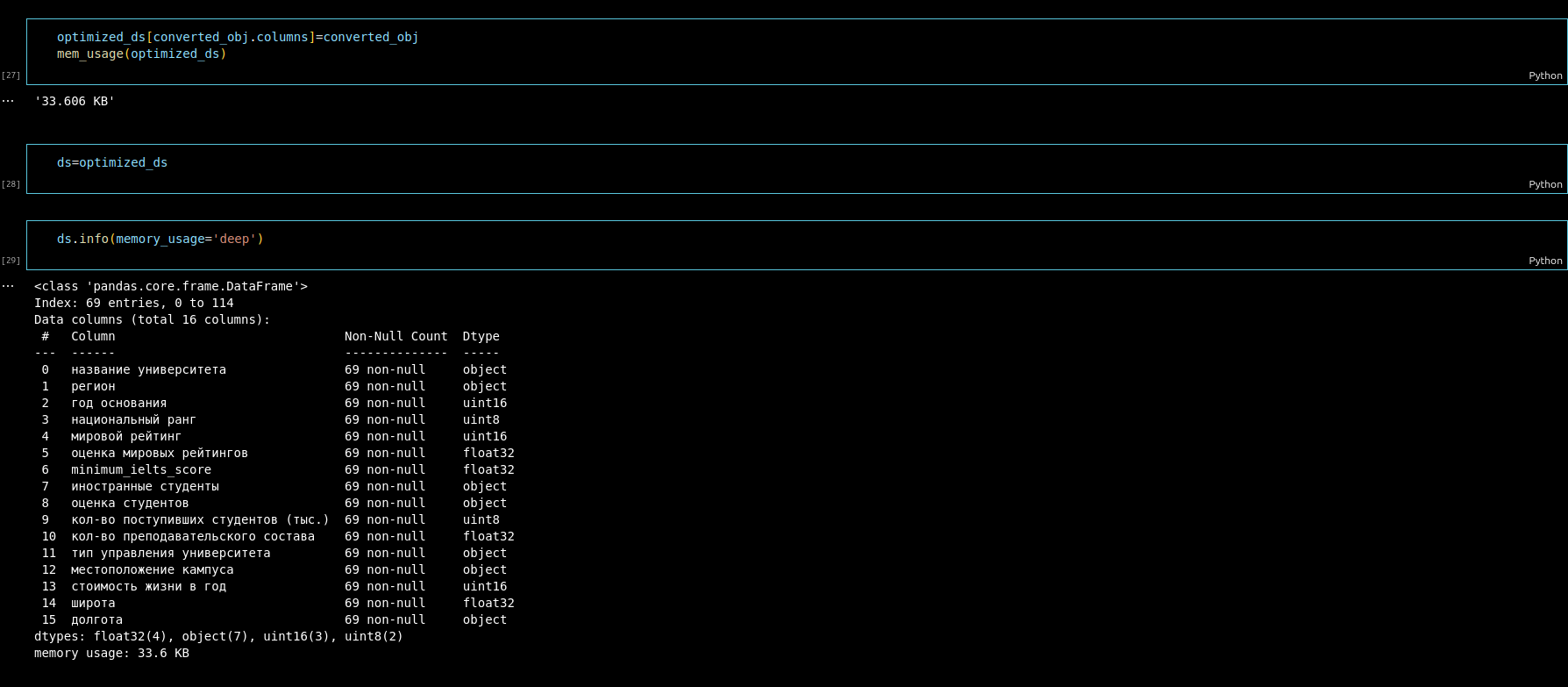


Создадим цикл, который перебирает все столбцы, хранящие данные типа object, выясняет, не превышает ли число уникальных значений в столбцах 50% и, если это так, преобразует их в тип category:



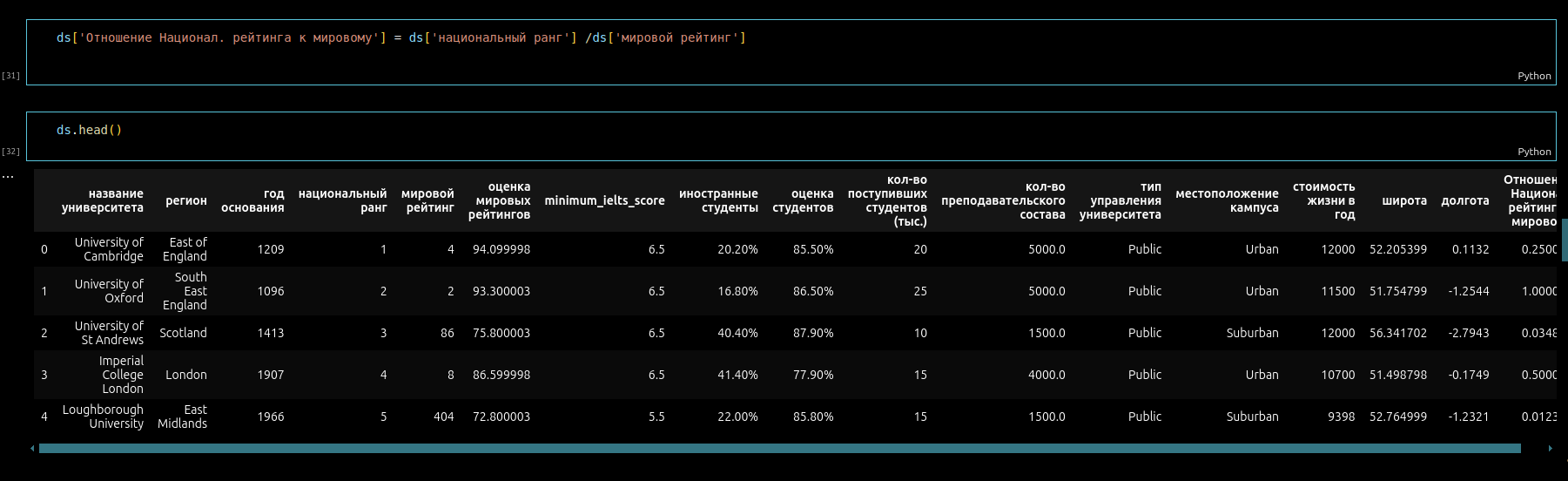
Теперь сравним то, что получилось после оптимизации, с тем, что было раньше:





**Агрегирование данных**

Добавим новые столбцы: мировой рейтинг и национальный ранг:



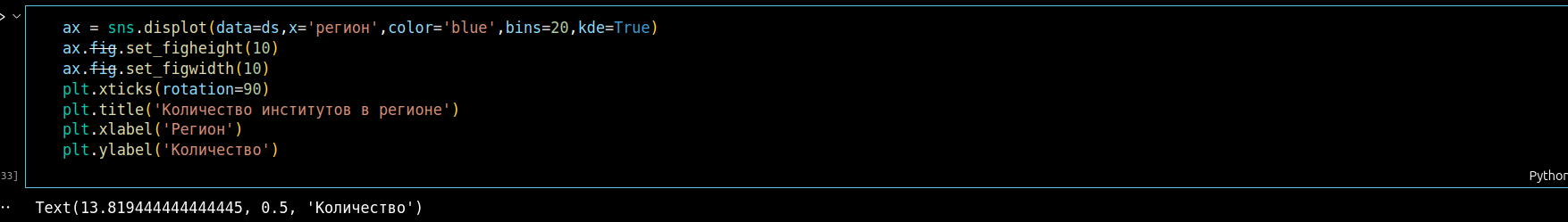
**Вывод:**

1. Была произведена оптимизация памяти и изменение типов данных для дальнейшего анализа [1].

# **Исследовательский анализ данных**

* 1. **Распределение университетов по регионам**

Есть различные университеты, такие как Кэмбридж, Оксфордский Университет, Академия в Лондоне и т.д. На рисунке представлен код, с помощью которого мы определяем количество университетов в каждом регионе.

****

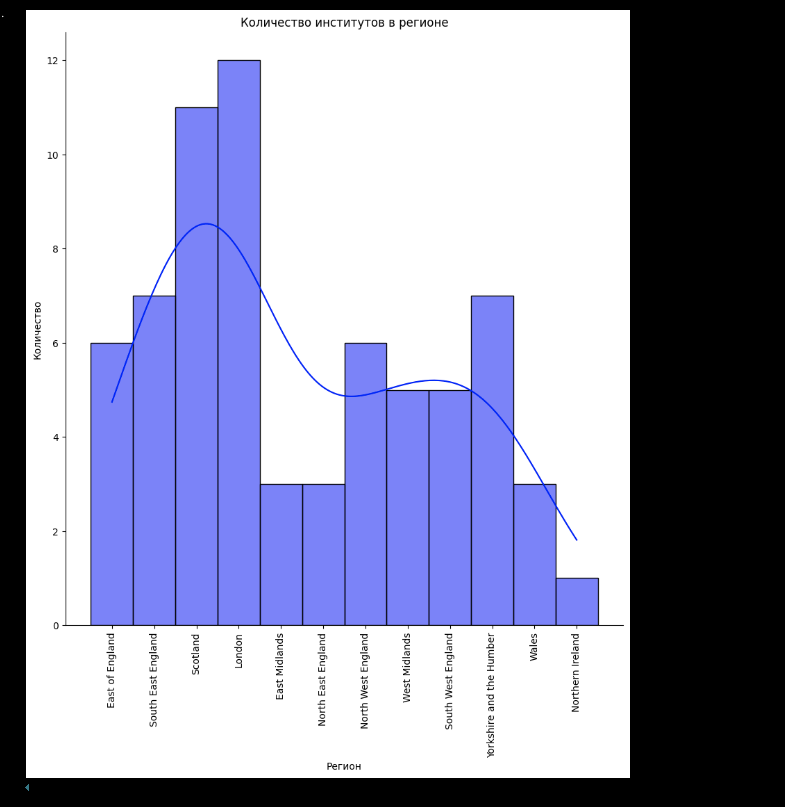
****

Рисунок 7.1 – Количество университетов в регионе

**Вывод:**

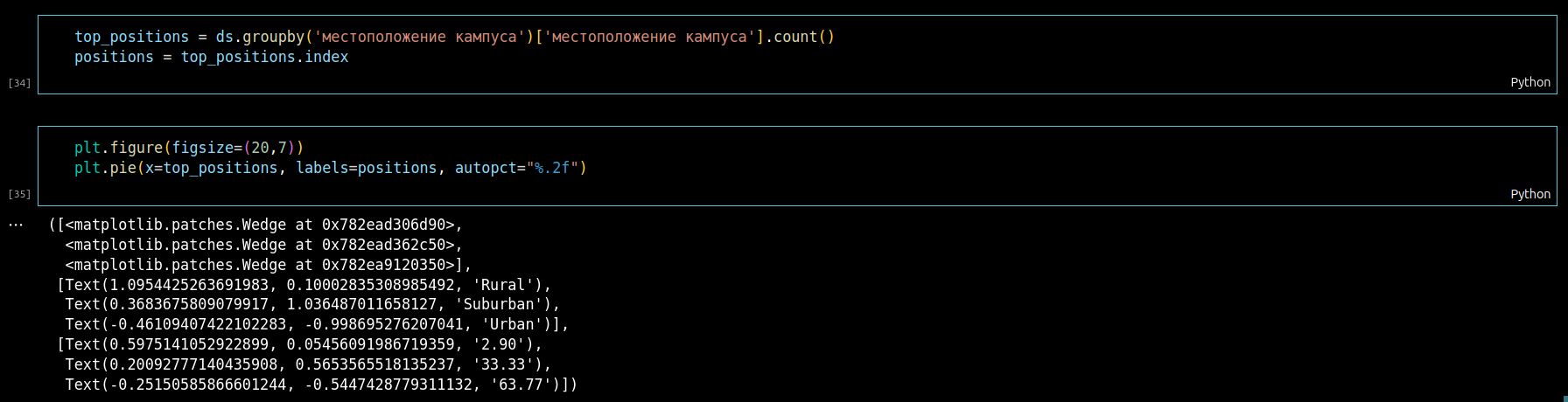
По графику на рисунке 7.1 можно заметить, что большинство университетов находится в Лондоне, это обусловлено хорошим качеством обучения и численностью населения, размером города [2].

Гипотеза 1 – подтверждена.

* 1. **Распределение университетов по местоположению кампуса**

Есть разные расположения местонахождений кампусов университетов, такие как городской, пригородный и деревенский.

Есть различные университеты, такие как Кэмбридж, Оксфордский Университет, Академия в Лондоне и т.д. На рисунке представлен код, с помощью которого мы определяем местоположения университетов.

****

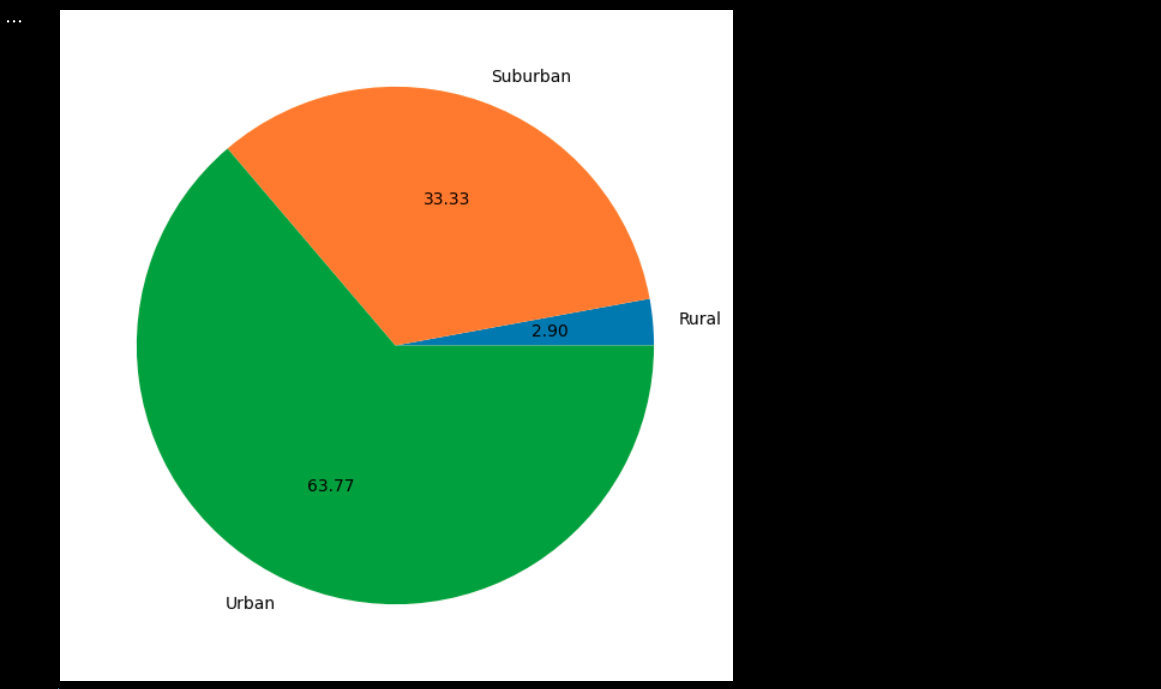
****

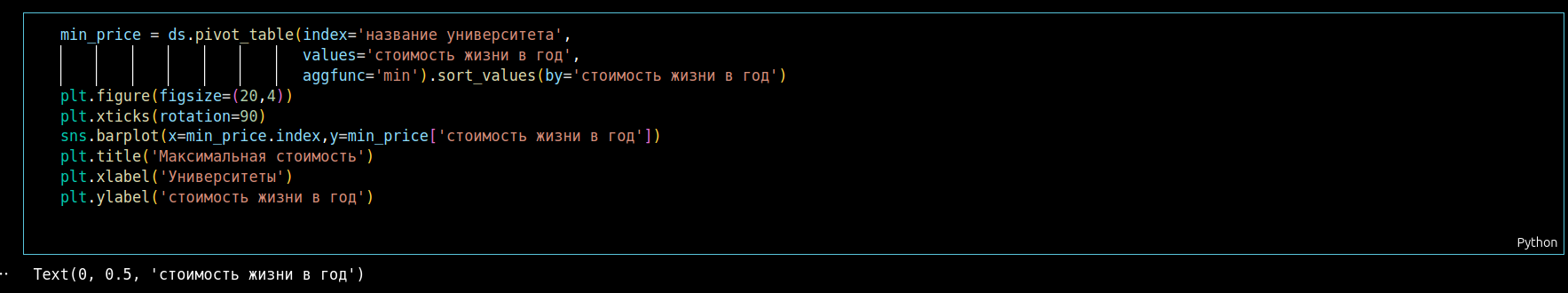
Рисунок 7.2 – Распределение университетов по местонахождению кампуса

**Вывод:**

Из графика на рисунке 7.2 видно, что больше всего кампусов находится в пределах города, а в деревнях хоть и есть университеты, но их мизерное количество [3].

Гипотеза 2 – доказана.

**Стоимость жизни в год**



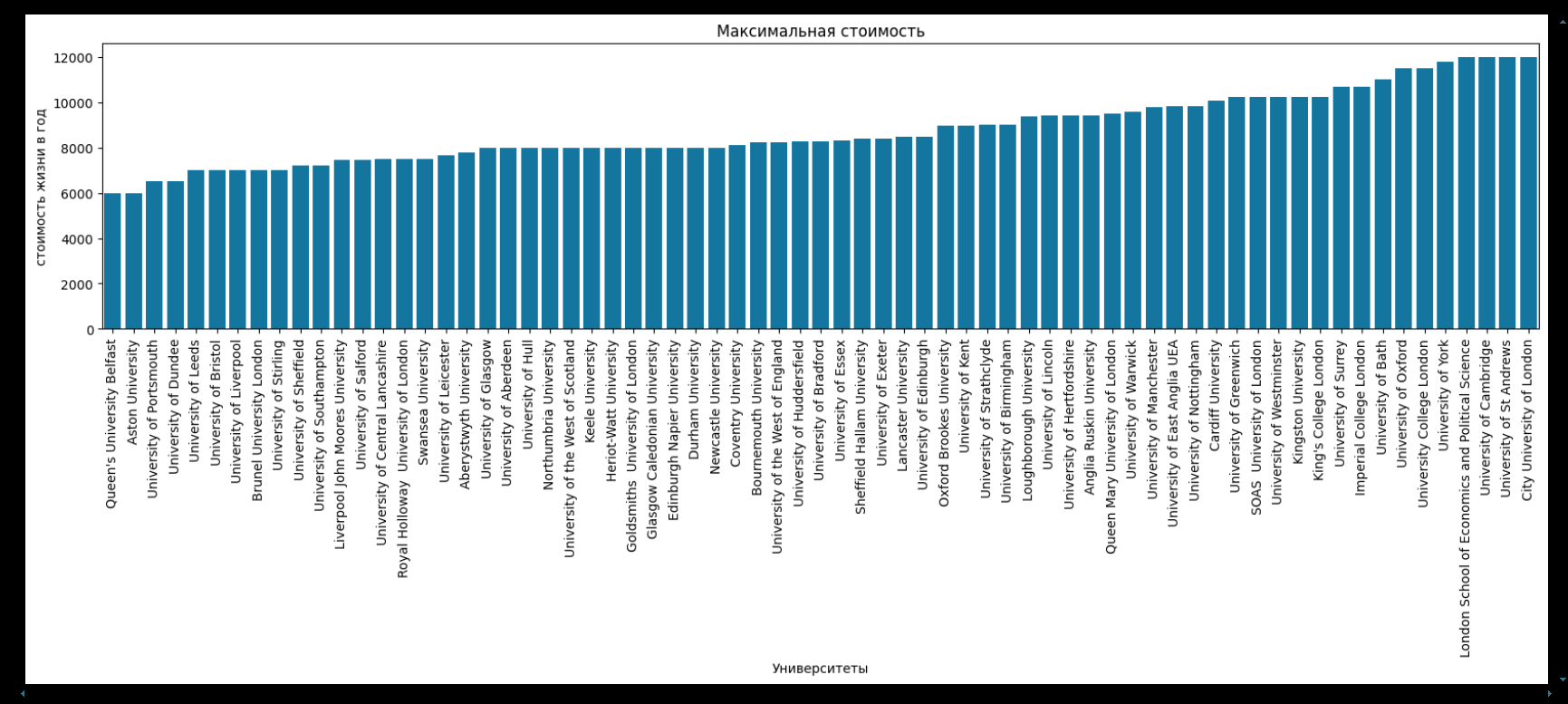


Рисунок 7.4 – стоимость жизни в год

**Вывод:**

Согласно диаграмме на рисунке 7.4 видно, что самая большая стоимость в университетах: Кэмбридж, University of St Andrews, City University of London. [5].

Будем считать, что гипотеза 4– доказана.

**Анализ корреляции**

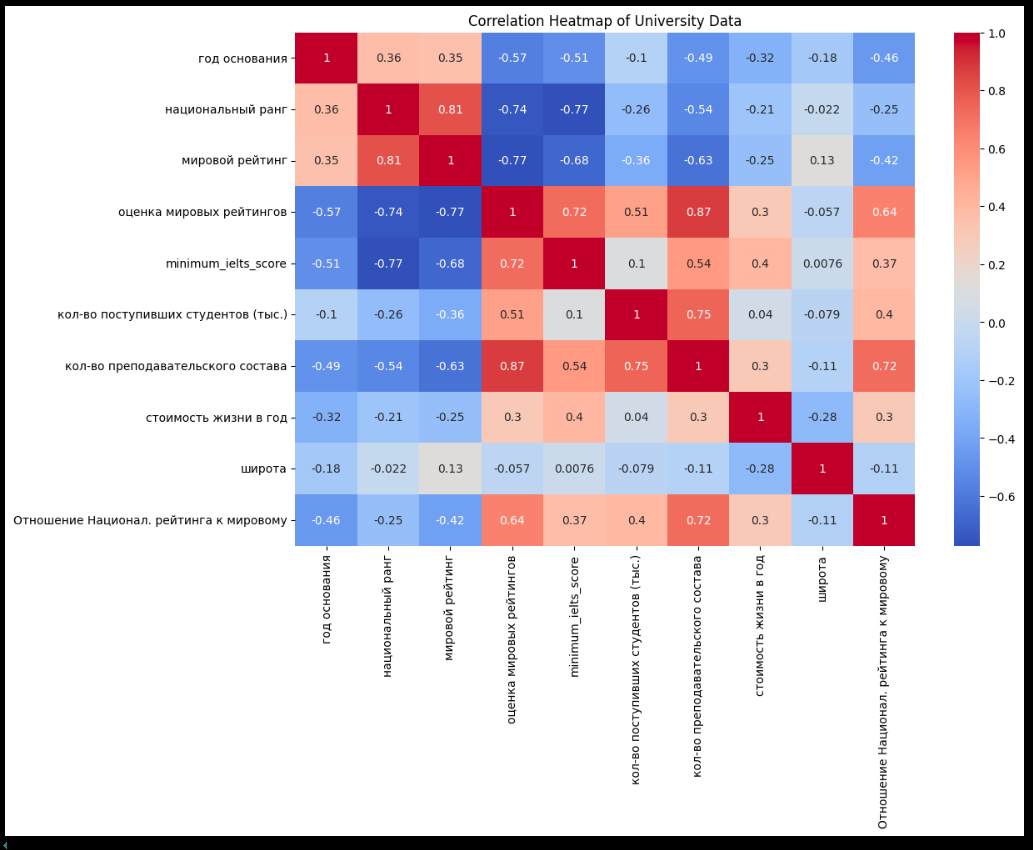
****

Рисунок 7.5 – Диаграмма корреляции

**Выводы корреляционного анализа:**

* 1. Чем выше оценка мировых рейтингов, тем выше баллы для поступления
  2. Чем больше преподавательского состава, тем выше оценка мировых рейтингов
  3. Чем выше мировой рейтинг, тем больше национальный ранг
  4. Стоимость жизни в год практически не зависит ни от каких факторов.

# **Общий вывод**

Перед анализом было оптимизированно использование памяти путем преобразования данных к более корректным типам. Также были добавлены необходимые столбцы при помощи агрегирования данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания в предметной области «Best Universities in the United Kingdom».

В ходе анализа было выявлено:

1. больше всего университетов в Лондоне;
2. большее количество университетов находится в пределах городской черты;
3. самый лучший университет Соединенного Королевства – Кэмбридж;
4. самая большая стоимость жизни в Кэмбридже, University of St Andrews, City University of London.

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ базы данных «Best Universities in the United Kingdom», выявлены важные статистические данные, которые могут помочь выпускникам выбрать университет для поступления [6].

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas» и «Seaborn» на базе языка программирования Python. Также мы пользовались «Jupyter Python» в нём мы писали код и анализировали диаграммы. В ходе работы были получены новые навыки в анализе данных и работы с выше перечисленными программами [7].

Оптимизация использования памяти, агрегирование данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания в предметной области. В ходе анализа выявлены лучшие университеты в Соединенном Королевстве, самые дорогие из них, а также более предпочитаемые регионы для жизни и обучения.

# **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](https://devpractice.ru/);
6. [https://coderoad.wiki](https://coderoad.wiki/);
7. [https://www.delftstack.com](https://www.delftstack.com/).