|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | **Министерство науки и высшего образования Российской Федерации**  Калужский филиал  федерального государственного бюджетного  образовательного учреждения высшего образования  ***«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»***  ***(КФ МГТУ им. Н.Э. Баумана)*** |

|  |  |
| --- | --- |
| **ФАКУЛЬТЕТ** | **ИУК «Информатика и управление»** |
| **КАФЕДРА** | **ИУК5 «Информатика и вычислительная техника»** |

**Лабораторная работа №1**

**«Инструментальные средства, подготовка эксперимента и анализ результатов»**

**ДИСЦИПЛИНА: «Проектирование программного обеспечения»**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил: студент гр. ИУК4-11М | |  |  | ( | Сафронов Н.С. | ) |
|  |  |  | (подпись) |  | (Ф.И.О.) |  |
| Проверил: | |  |  | ( | Потапов А.Е. | ) |
|  |  |  | (подпись) |  | (Ф.И.О.) |  |

|  |  |
| --- | --- |
| Дата сдачи (защиты):  Результаты сдачи (защиты): | |
|  | - Балльная оценка:  - Оценка: |

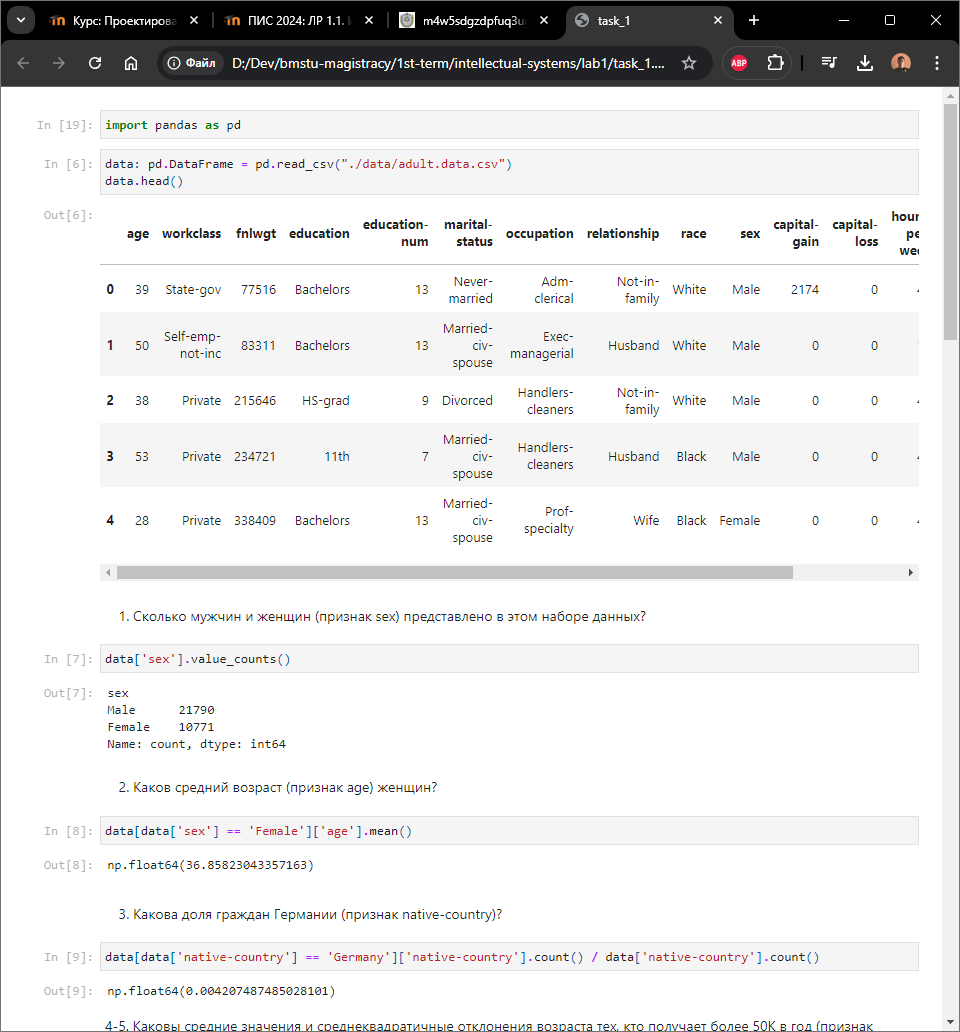
Калуга, 2024

**Цель работы:** формирование практических навыков подготовки и настройки инструментальной среды анализа данных, умений подготовить эксперимент, представить и проанализировать полученные данные.

**Задачи:** подготовить и настроить среду исполнения Python, выполнить предобработку данных эксперимента, реализовать простейшие алгоритмы анализа данных (получение статистических характеристик), реализовать адекватное графическое представление результатов**.**

**Результаты выполнения работы**

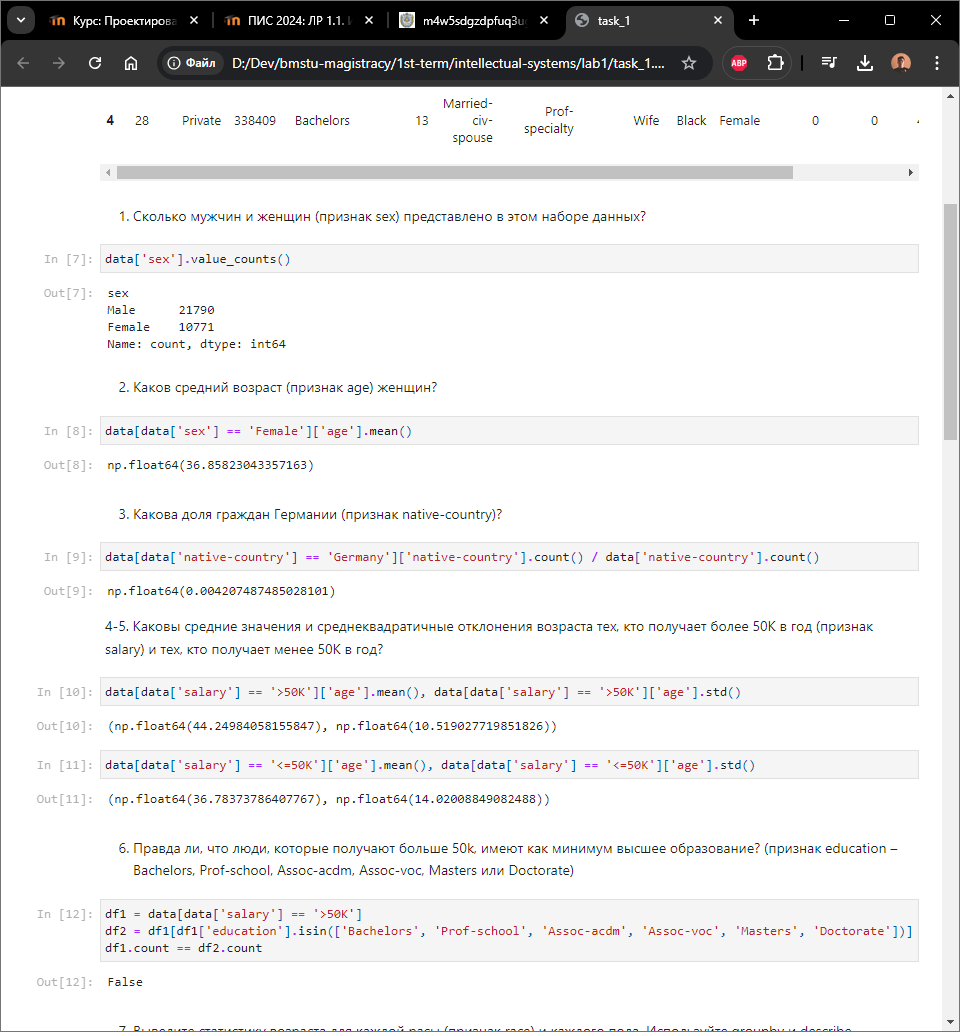
**Задание 1**



**Рисунок 1** – Результат загрузки данных в pandas

**Задача 1**

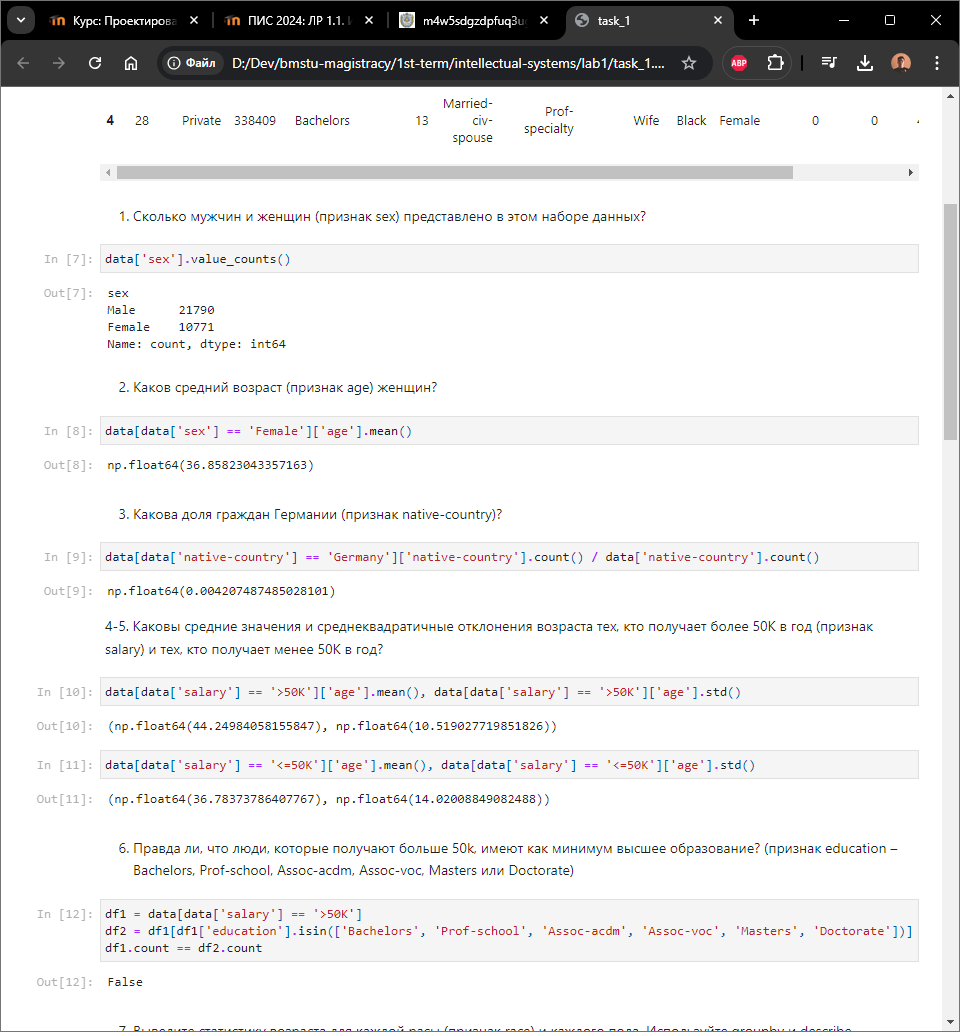
Сколько мужчин и женщин (признак sex) представлено в этом наборе данных?



**Рисунок 2** – Результат получения количества мужчин и женщин в выборке

**Задача 2**

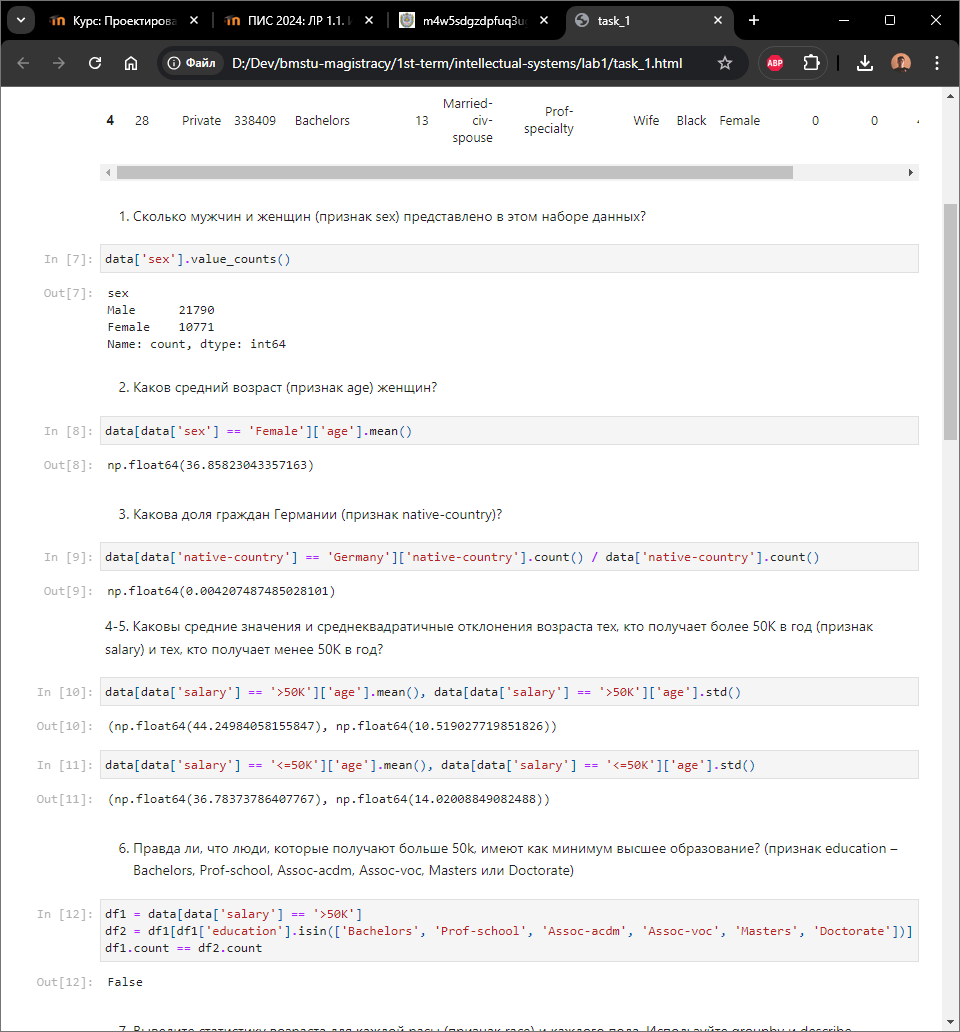
Каков средний возраст (признак age) женщин?



**Рисунок 3** – Вычисленный средний возраст женщин в выборке

**Задача 3**

Какова доля граждан Германии (признак native-country)?

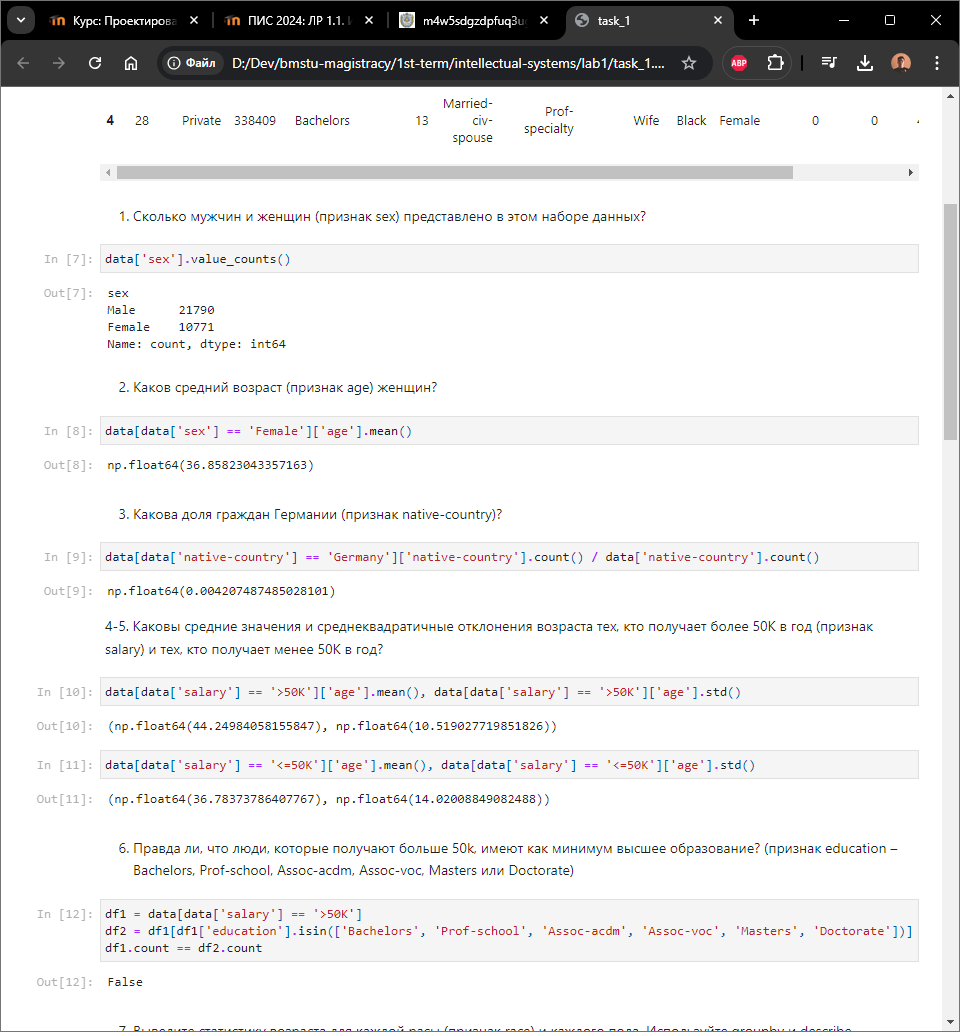


**Рисунок 4** – Доля граждан Германии в выборке

Таким образом, доля граждан Германии – 0.0042 или 0.42%.

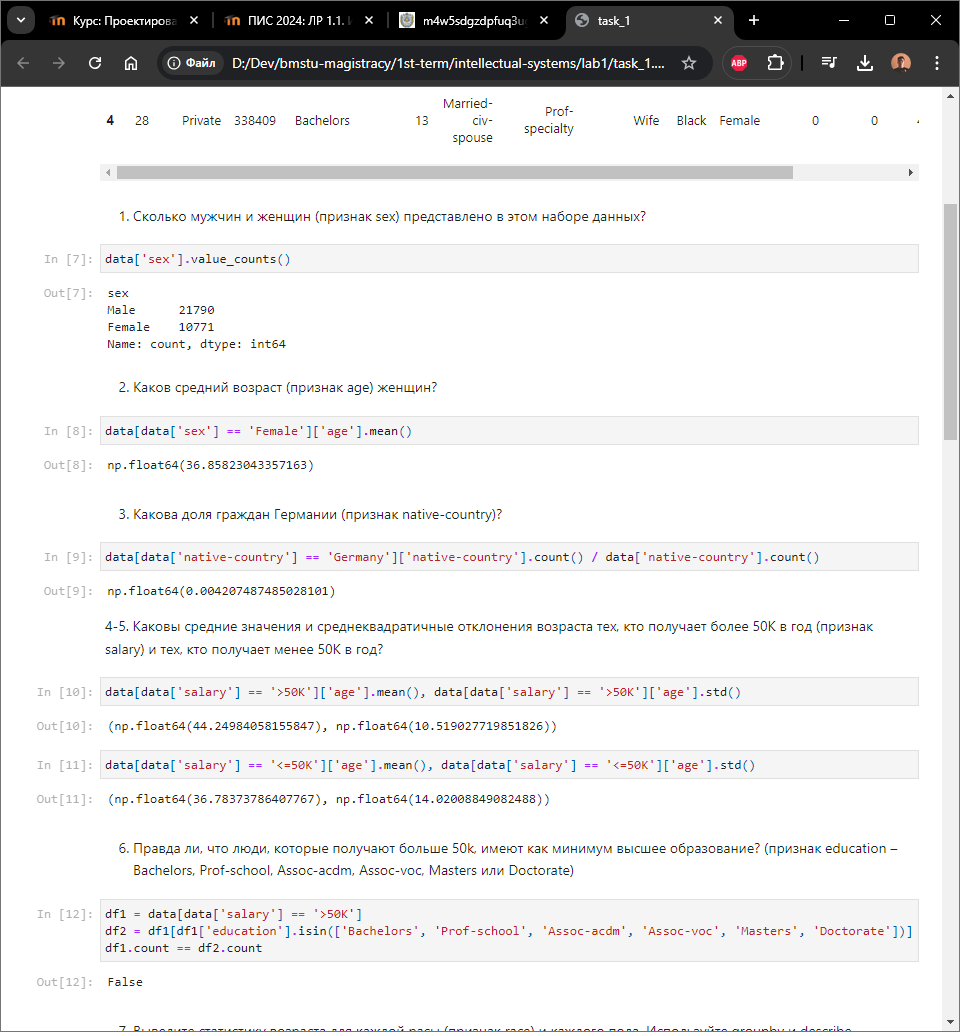
**Задачи 4-5**

Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50K в год (признак salary) и тех, кто получает менее 50K в год?



**Рисунок 5** – Среднее значение и среднеквадратичное отклонение возраста тех, кто получает более 50К в год

Таким образом, среднее значение возраста тех, кто получает более 50К в год, - 44.25; среднеквадратичные отклонения – 10.52.

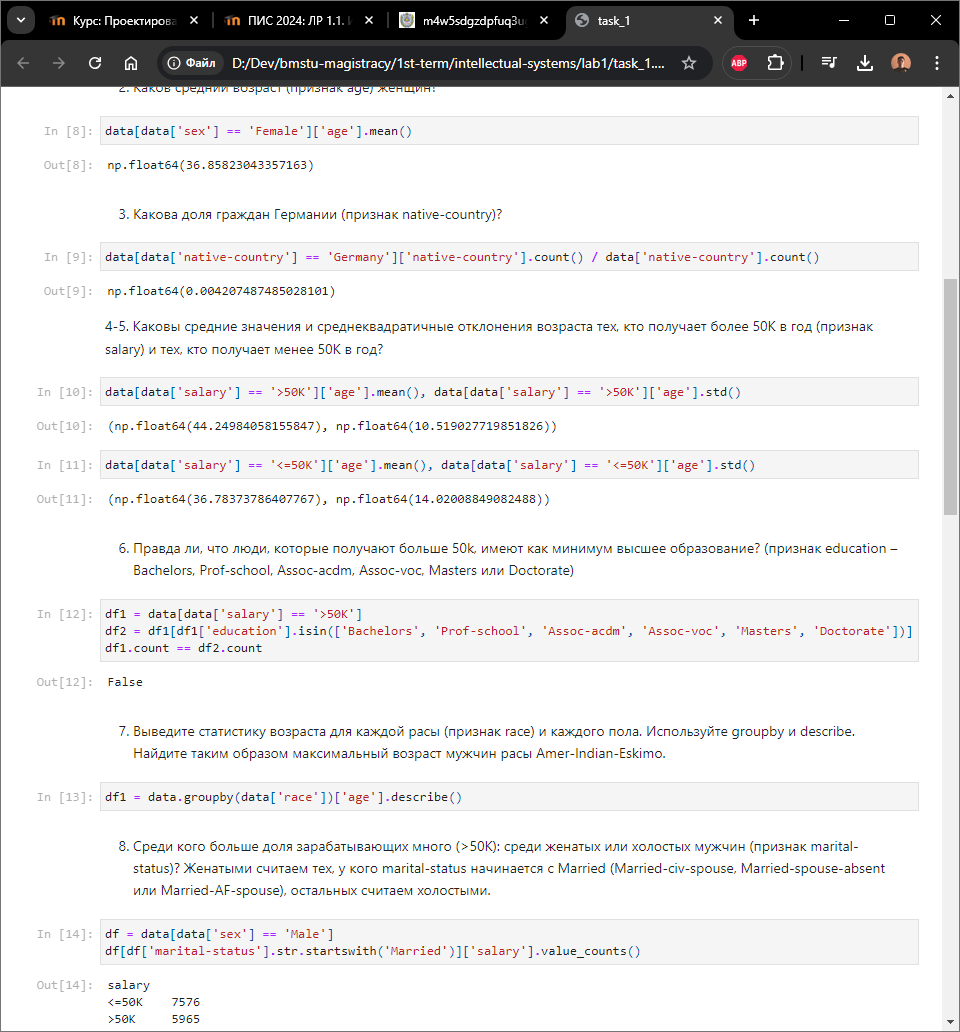


**Рисунок 6** – Среднее значение и среднеквадратичное отклонение возраста тех, кто получает менее либо 50К в год

Таким образом, среднее значение возраста тех, кто получает более 50К в год, - 36.78; среднеквадратичные отклонения – 14.02.

**Задача 6**

Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)

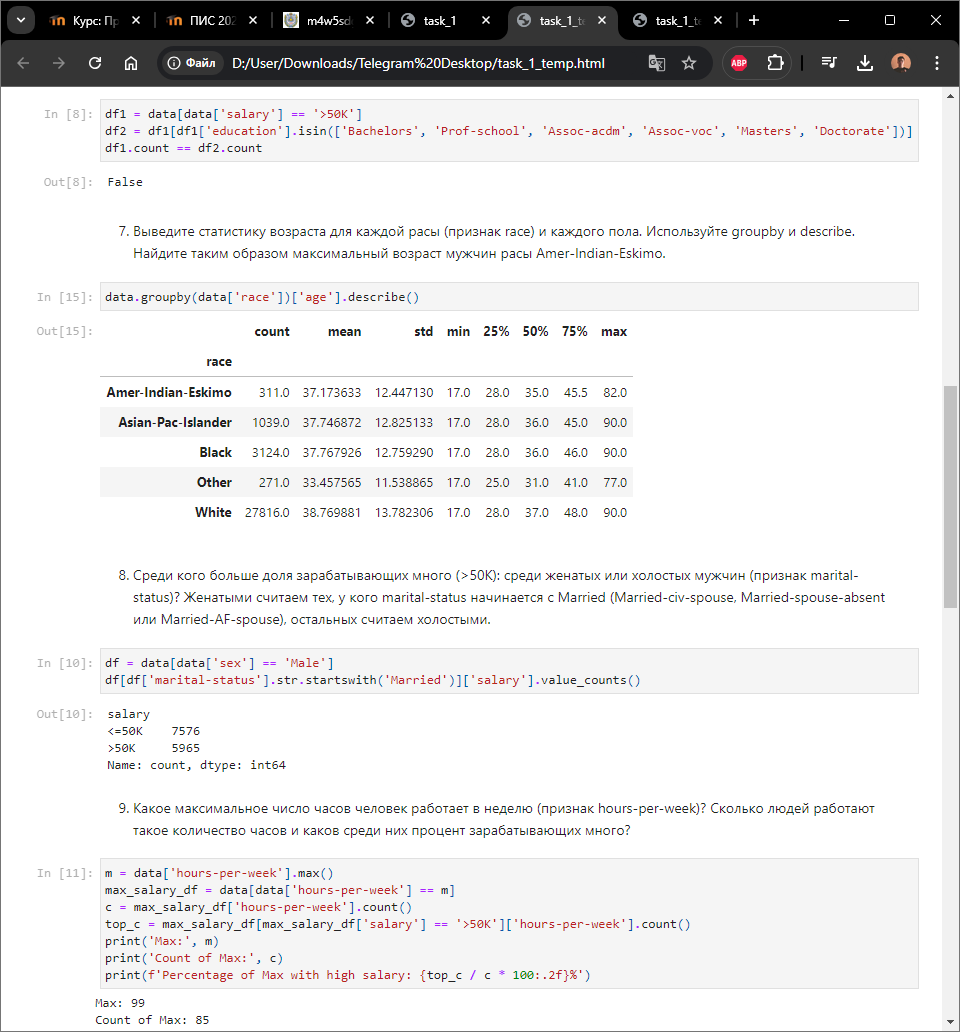


**Рисунок 7** – Результат сравнения числа тех, кто имеет и не имеет высшее образование, среди зарабатывающих более 50К

Таким образом, утверждение не является правдой, так как число тех, кто имеет и не имеет высшее образование, среди зарабатывающих более 50К не является равным.

**Задача 7**

Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.

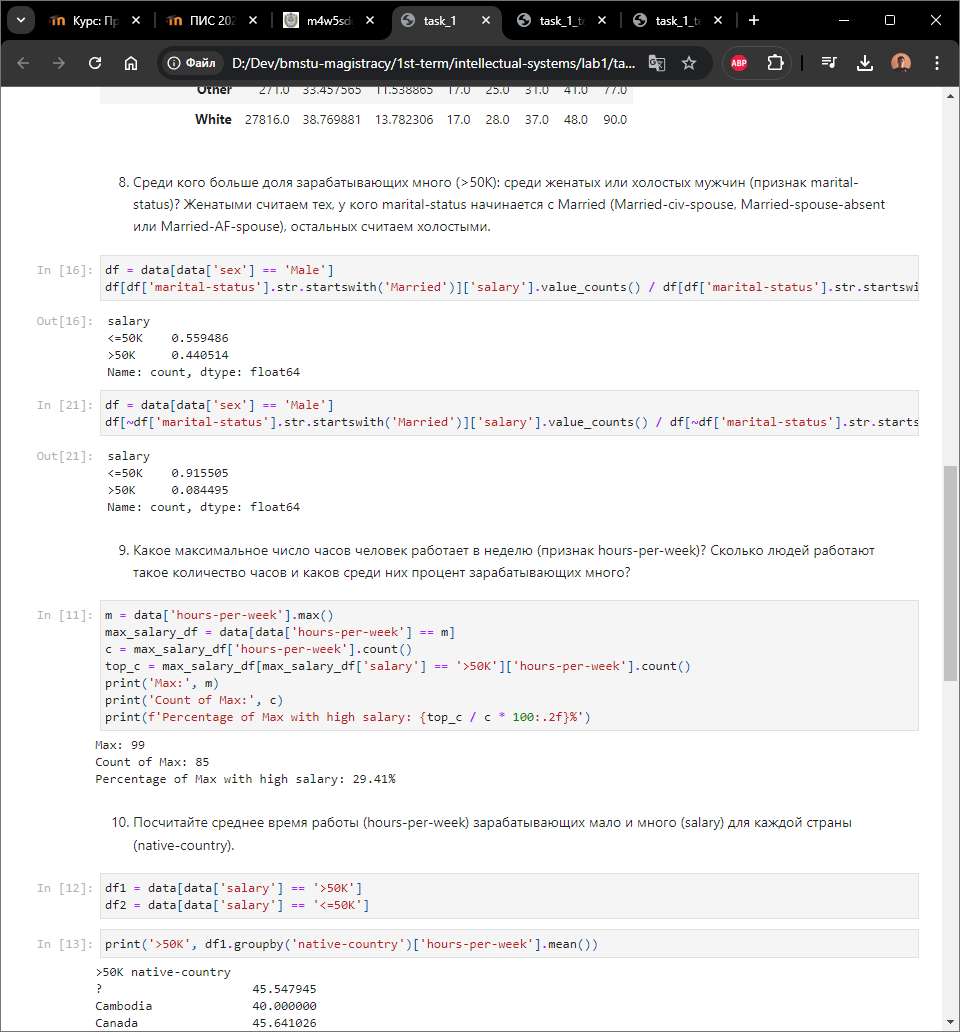


**Рисунок 8** – Статистика возраста для каждой расы

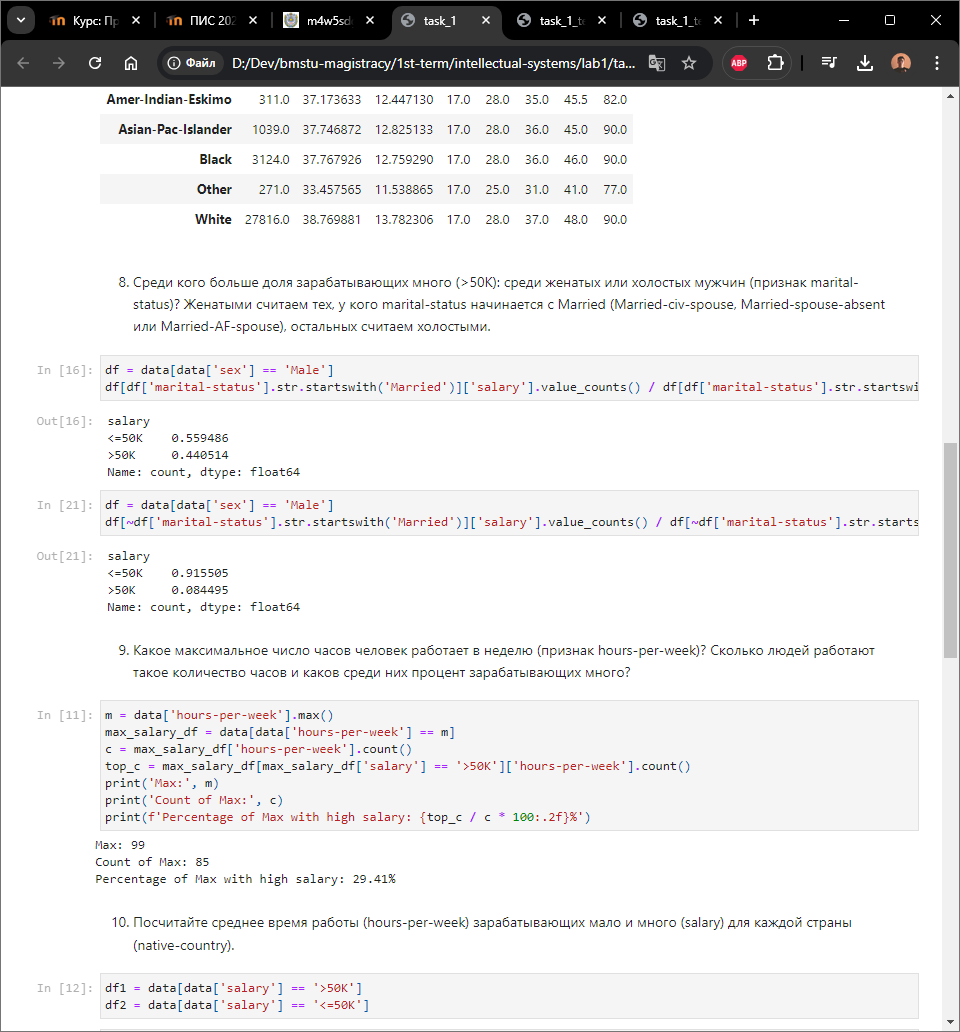
Исходя из данных таблицы, максимальный возраст мужчин расы Amer-Indian-Eskimo – 82 года.

**Задача 8**

Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.



**Рисунок 9** – Доля зарабатывающих много и мало среди женатых

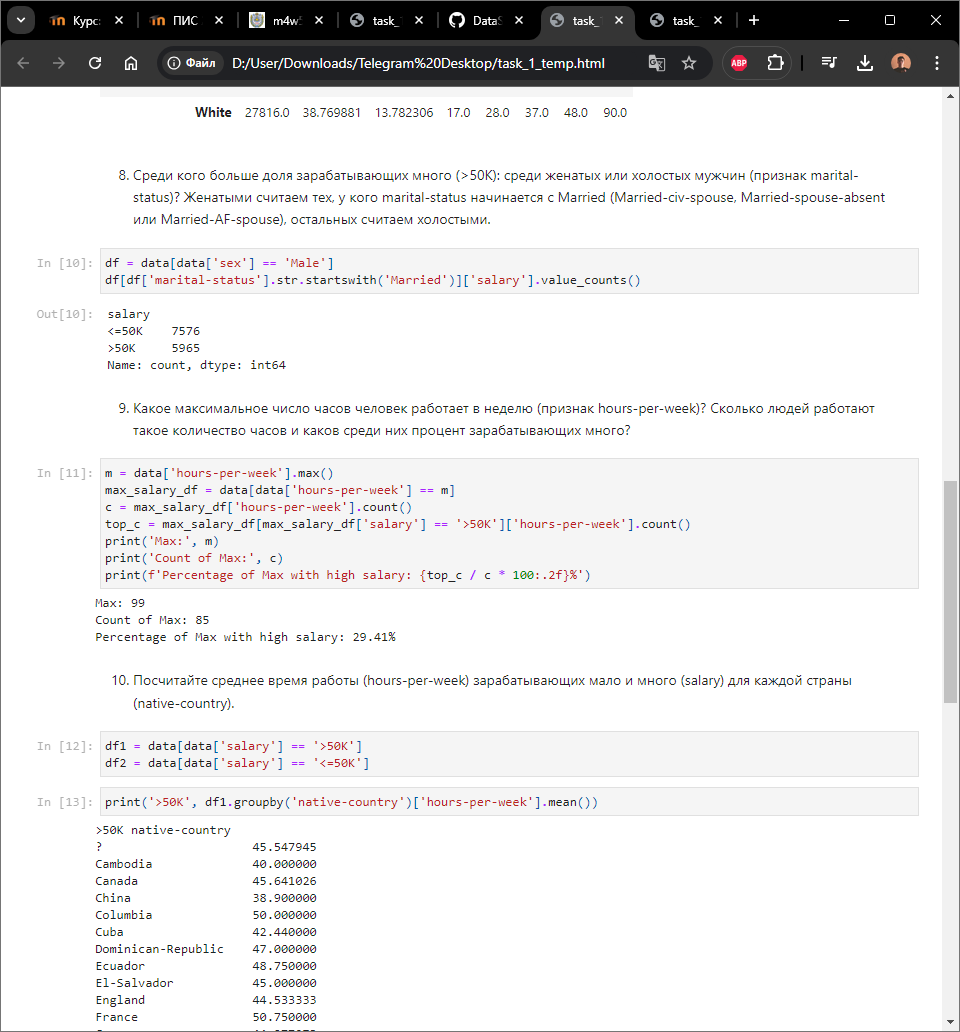


**Рисунок 10** – Доля зарабатывающих много и мало среди неженатых

Таким образом, доля зарабатывающих много больше у женатых мужчин.

**Задача 9**

Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?



**Рисунок 11** – Расчёты для времени работы в неделю

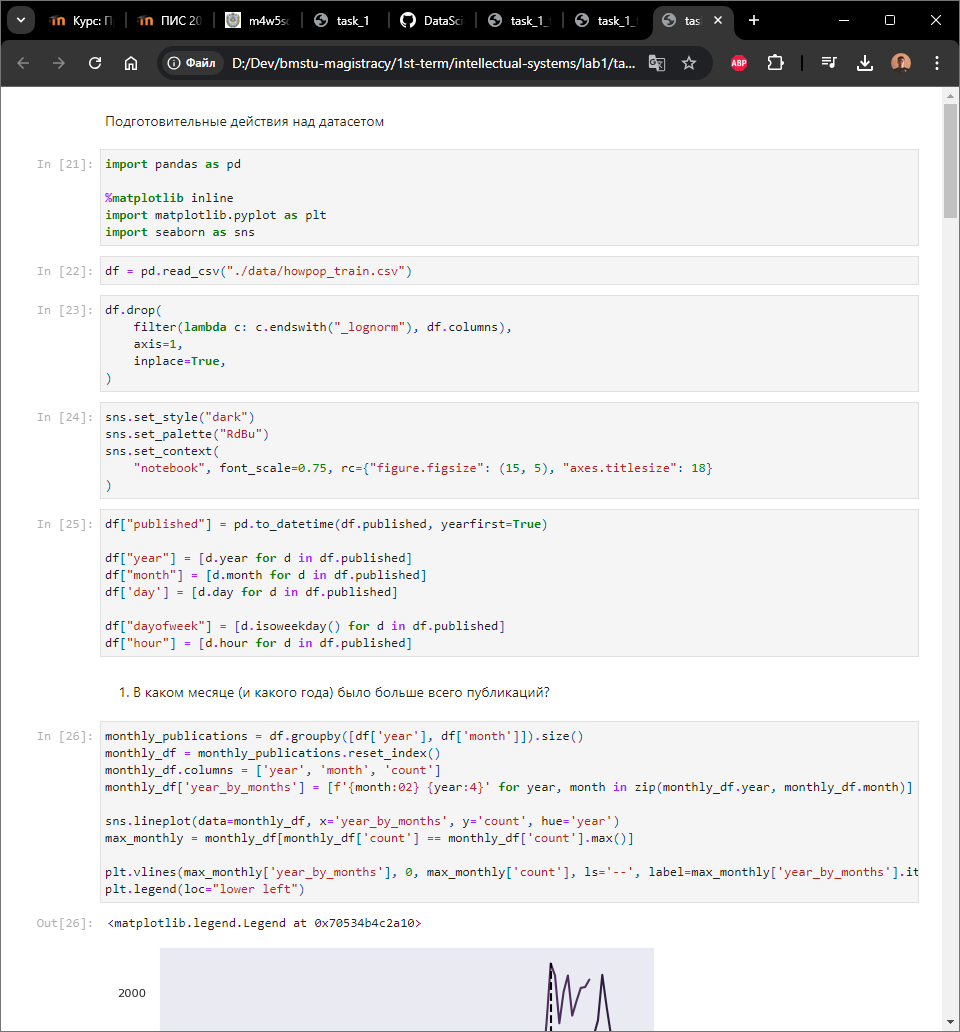
Таким образом, максимальное число часов человек работает в неделю равно 99 часам; 85 человек работают такое количество часов и 29.41% из них зарабатывают много.

**Задача 10**

Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).

|  |  |
| --- | --- |
|  |  |
| **Рисунок 12** – Среднее время работы зарабатывающих много по странам | **Рисунок 13** – Среднее время работы зарабатывающих мало по странам |

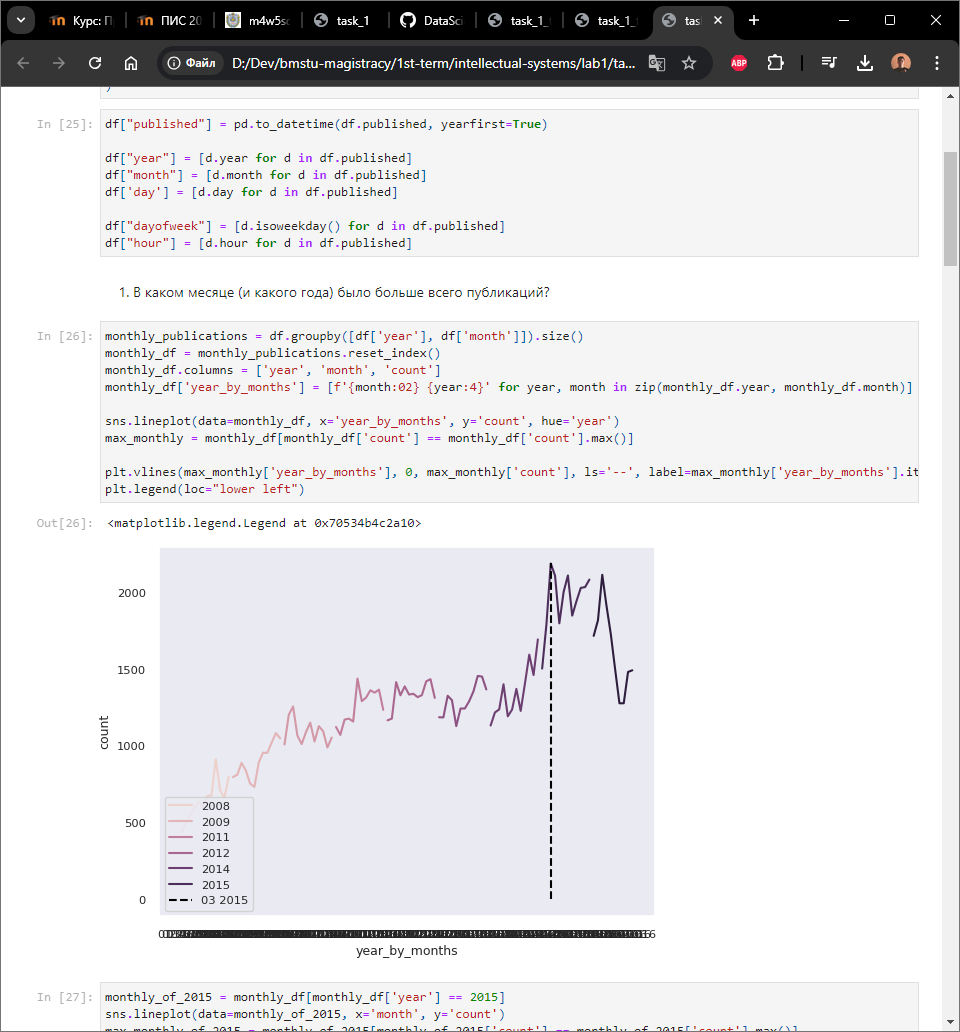
**Задание 2**



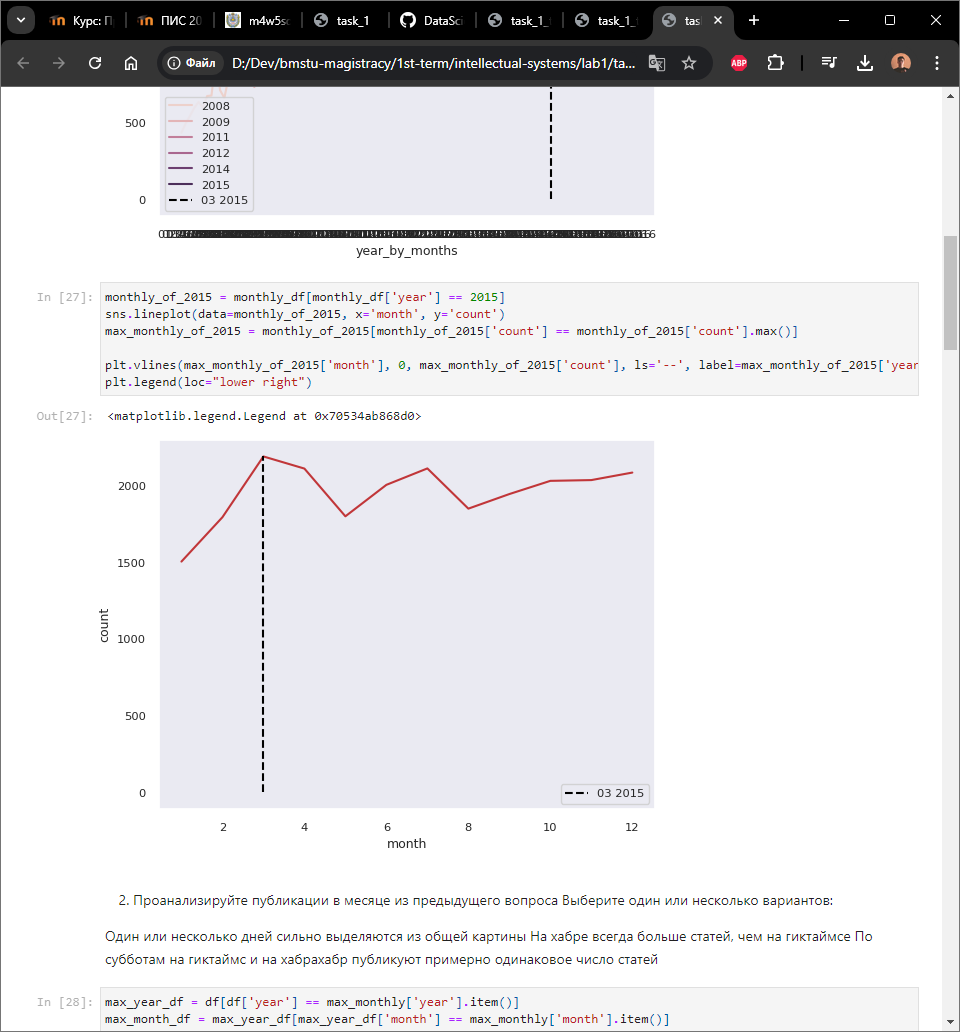
**Рисунок 14 –** Подготовительные действия над датасетом

**Задача 1**

В каком месяце (и какого года) было больше всего публикаций?



**Рисунок 15 –** Графическое отображение максимума на графике числа публикаций в месяц



**Рисунок 16 –** Число публикаций в 2015 году по месяцам

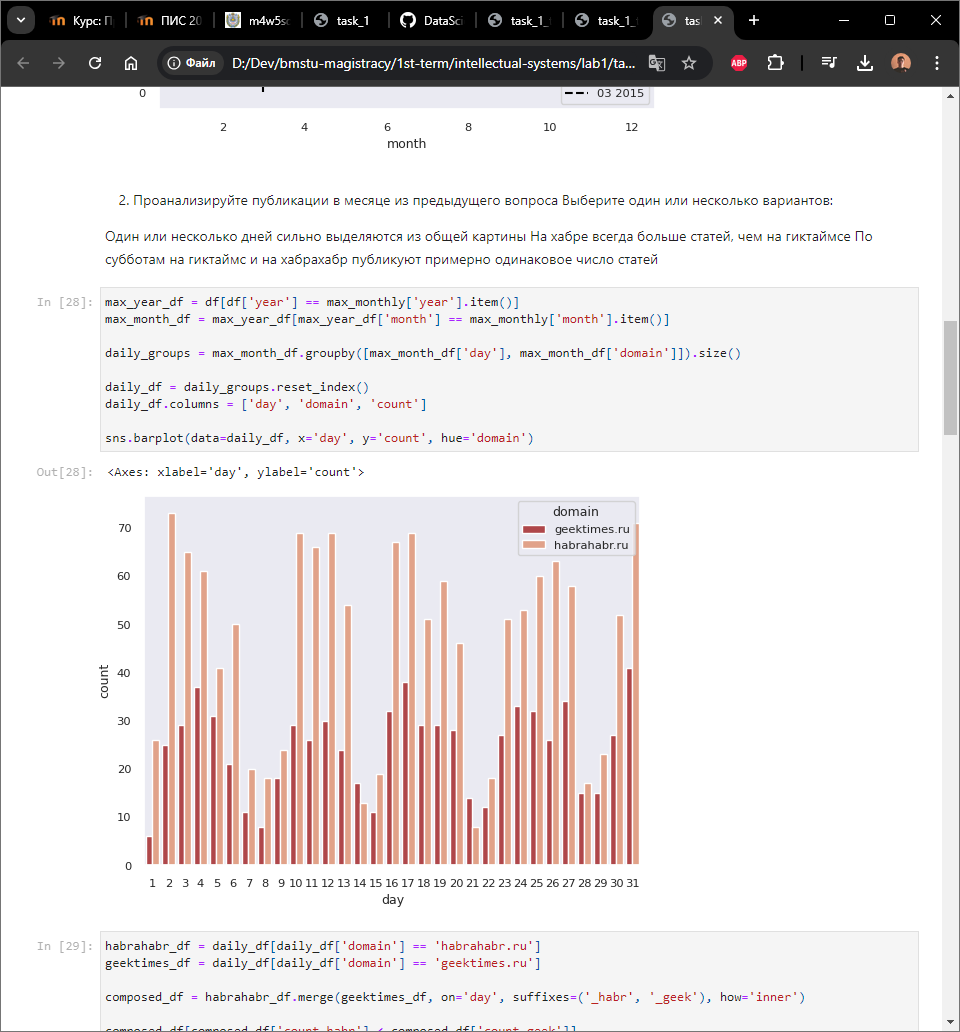
Получаем, что наибольшее число публикаций было в марте 2015 года.

**Задача 2**

Проанализируйте публикации в месяце из предыдущего вопроса.

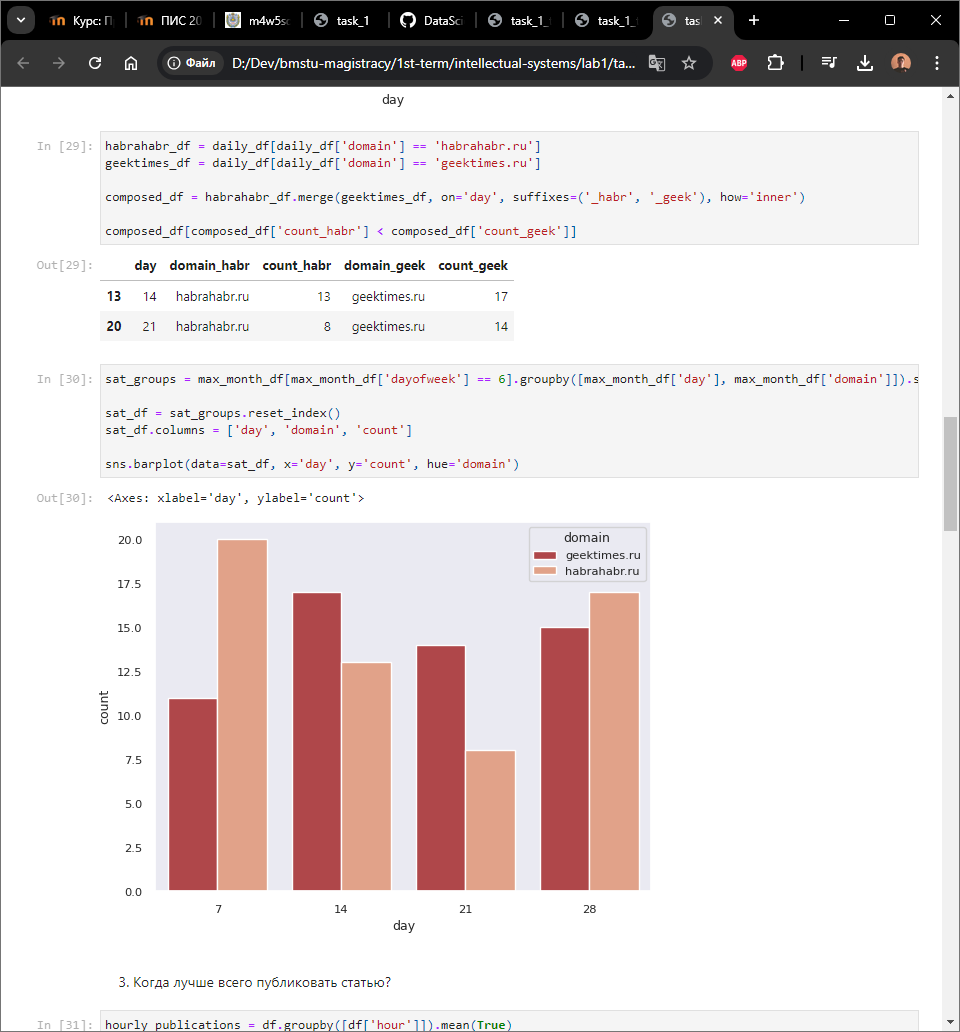
Выберите один или несколько вариантов:

* Один или несколько дней сильно выделяются из общей картины.
* На хабре всегда больше статей, чем на гиктаймсе.
* По субботам на гиктаймс и на хабрахабр публикуют примерно одинаковое число статей.



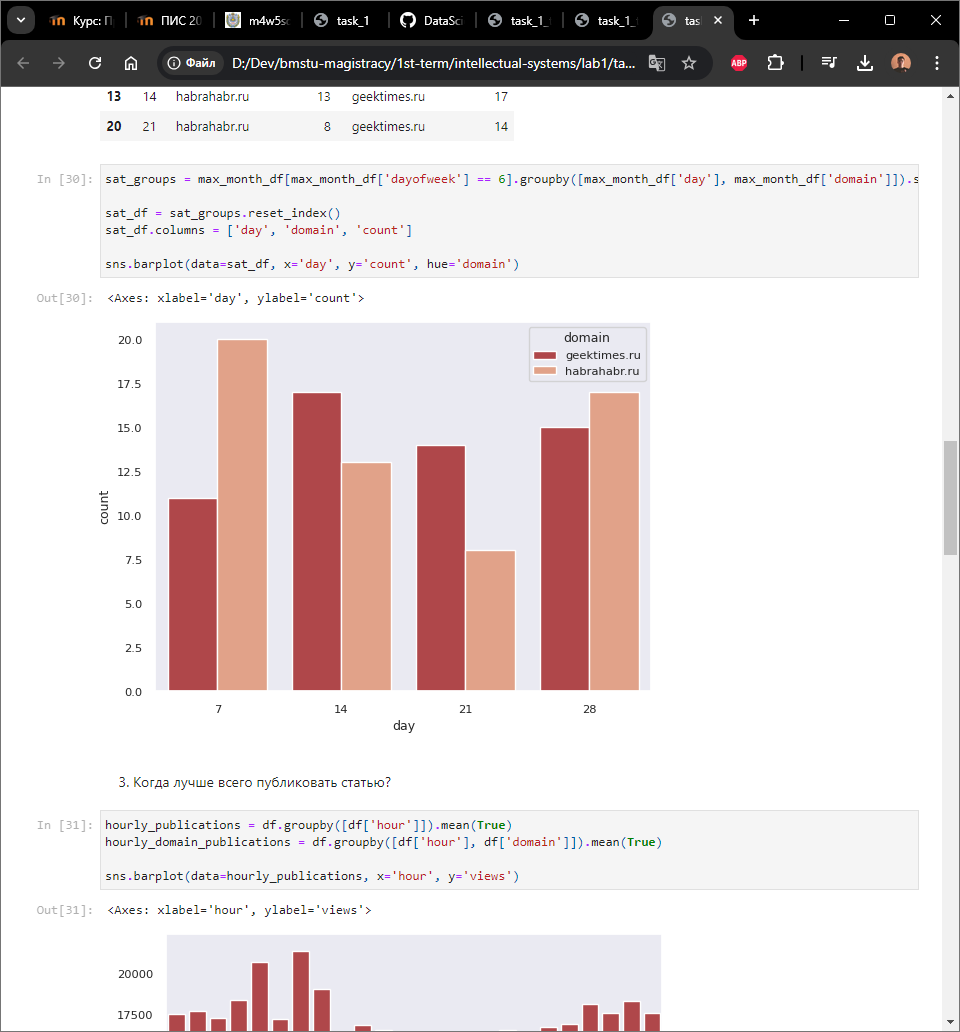
**Рисунок 17 –** Число публикаций в марте 2015 году по дням на разных доменах

Таким образом, график цикличен, а следовательно, ни один из дней не выделяется из общей картины.



**Рисунок 18 –** Дни, когда на гиктаймс было больше статей, чем на хабре

Таким образом, на хабре не всегда больше статей, чем на гиктаймсе.



**Рисунок 19 –** Публикации в субботу на доменах

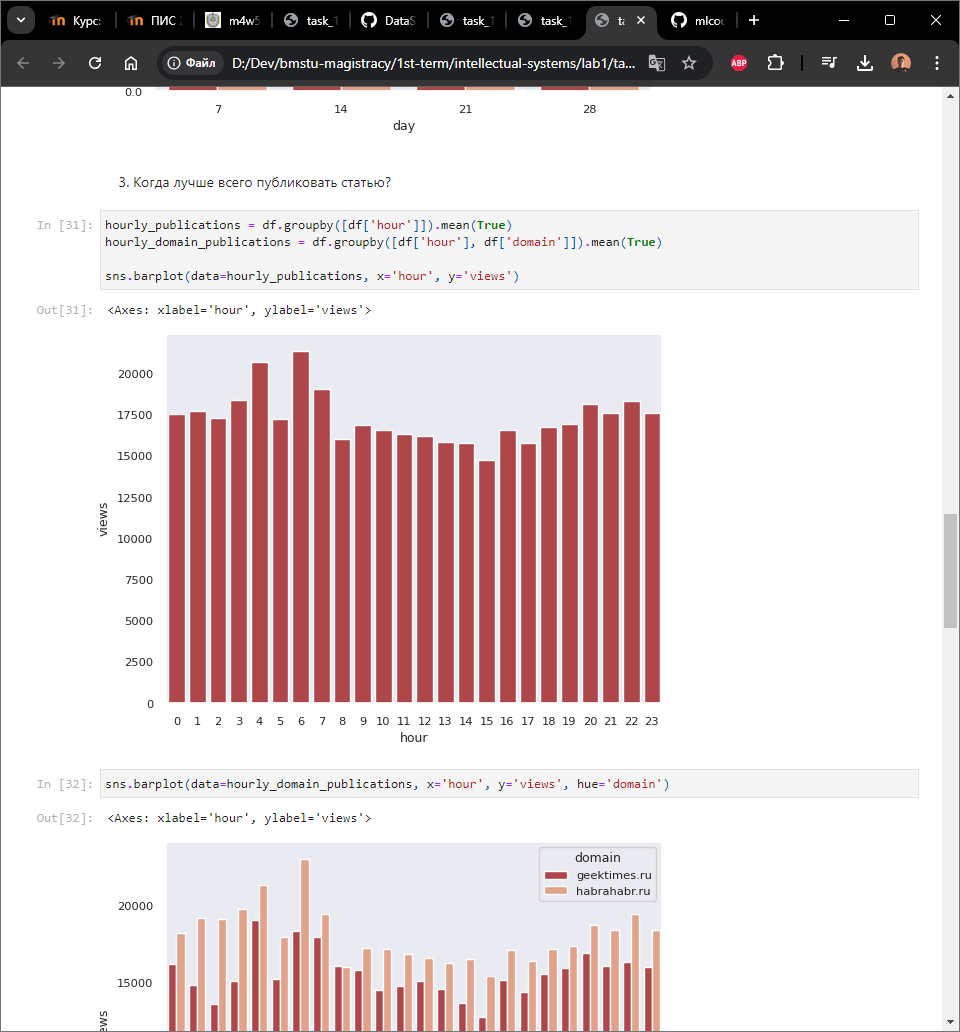
Таким образом, в субботу действительно на обоих доменах приблизительно одинаковое число статей.

Выбираем только пункт *«По субботам на гиктаймс и на хабрахабр публикуют примерно одинаковое число статей».*

**Задача 3**

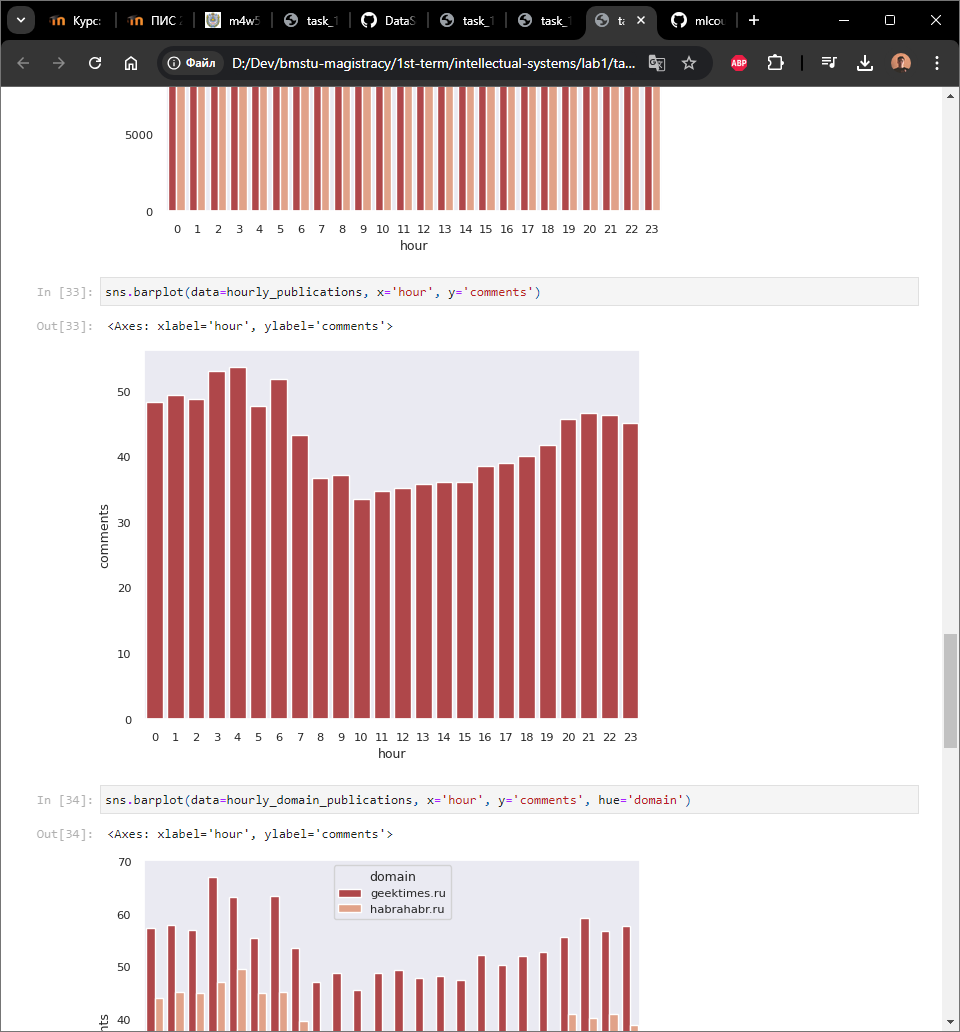
Когда лучше всего публиковать статью?

* Больше всего просмотров набирают статьи, опубликованные в 12 часов дня.
* У опубликованных в 10 утра постов больше всего комментариев.
* Больше всего просмотров набирают статьи, опубликованные в 6 часов утра.
* Максимальное число комментариев на гиктаймсе набрала статья, опубликованная в 9 часов вечера.
* На хабре дневные статьи комментируют чаще, чем вечерние.



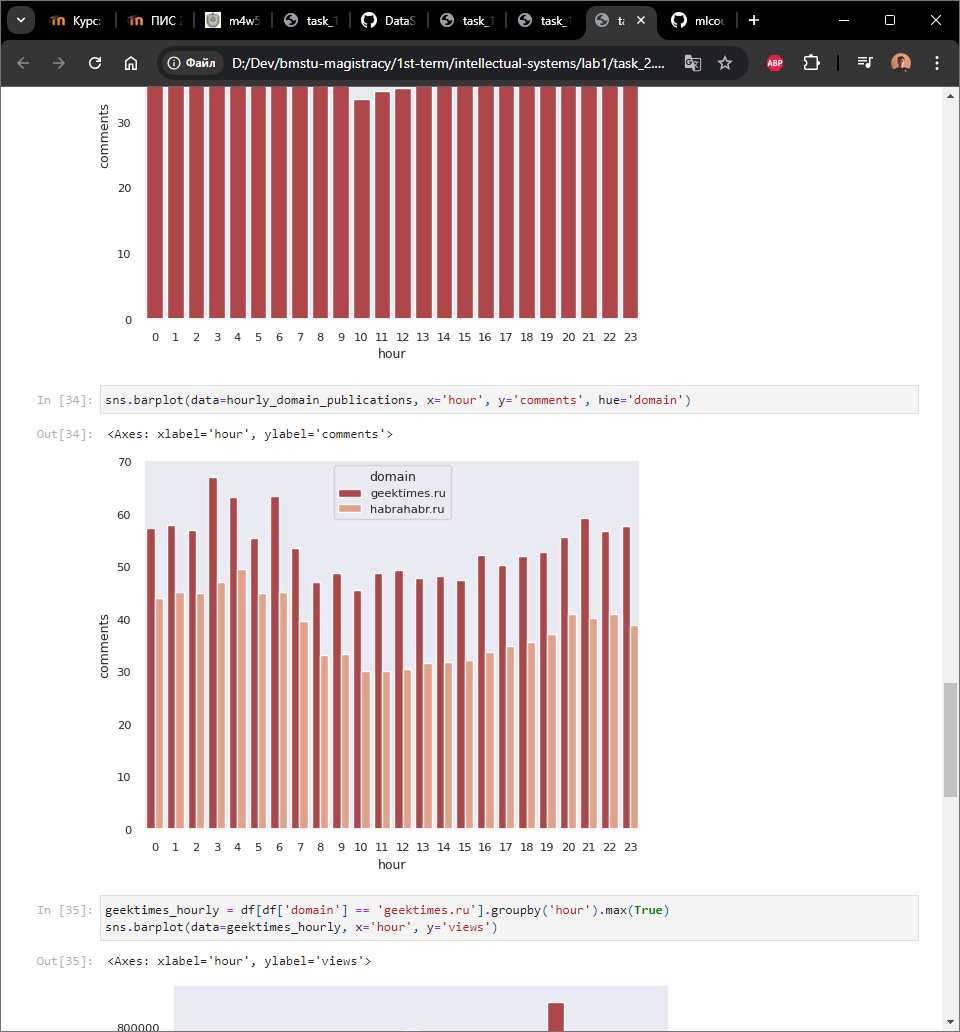
**Рисунок 20 –** Просмотры статей относительно времени

Таким образом, больше всего просмотров набирают статьи, опубликованные в 6 часов утра, а не в 12 часов дня.



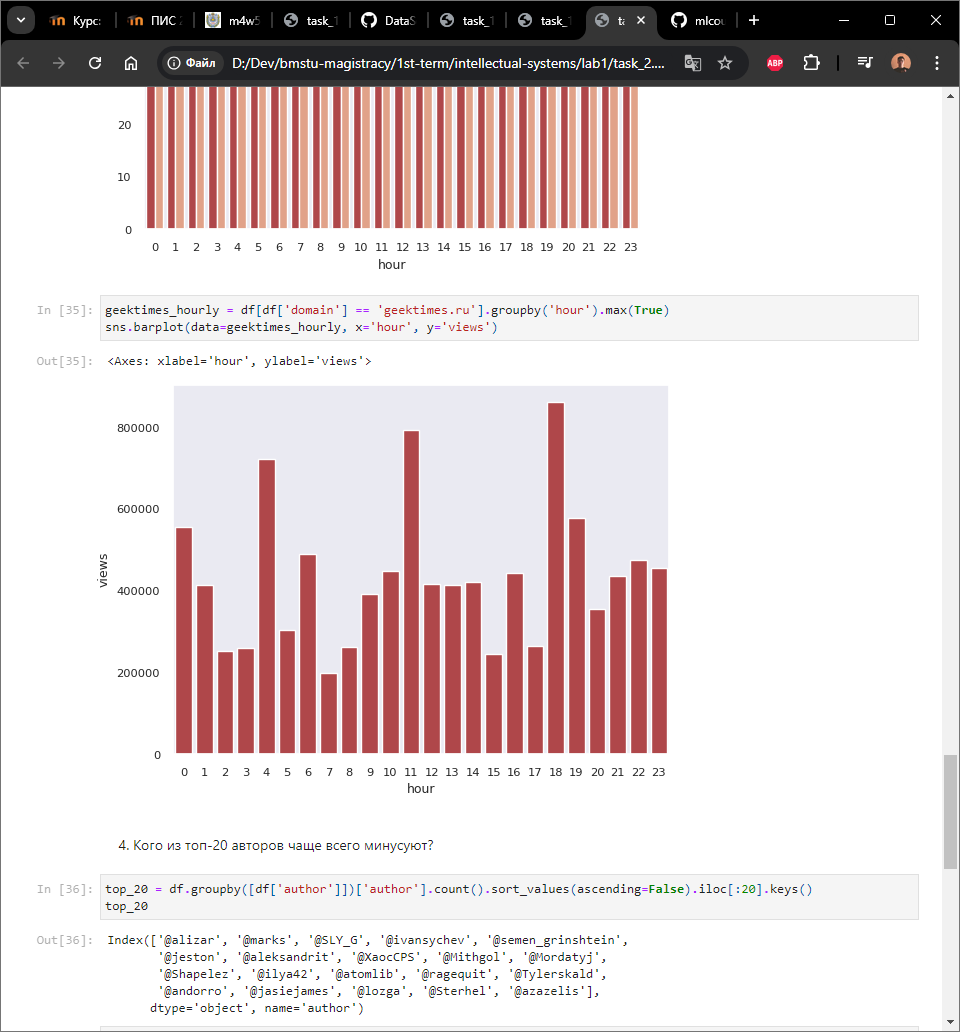
**Рисунок 21 –** Комментарии статей относительно времени

Таким образом, тезис «У опубликованных в 10 утра постов больше всего комментариев» опровергается.



**Рисунок 22 –** Комментарии статей относительно времени по доменам

Таким образом, тезис «На хабре дневные статьи комментируют чаще, чем вечерние» опровергается.



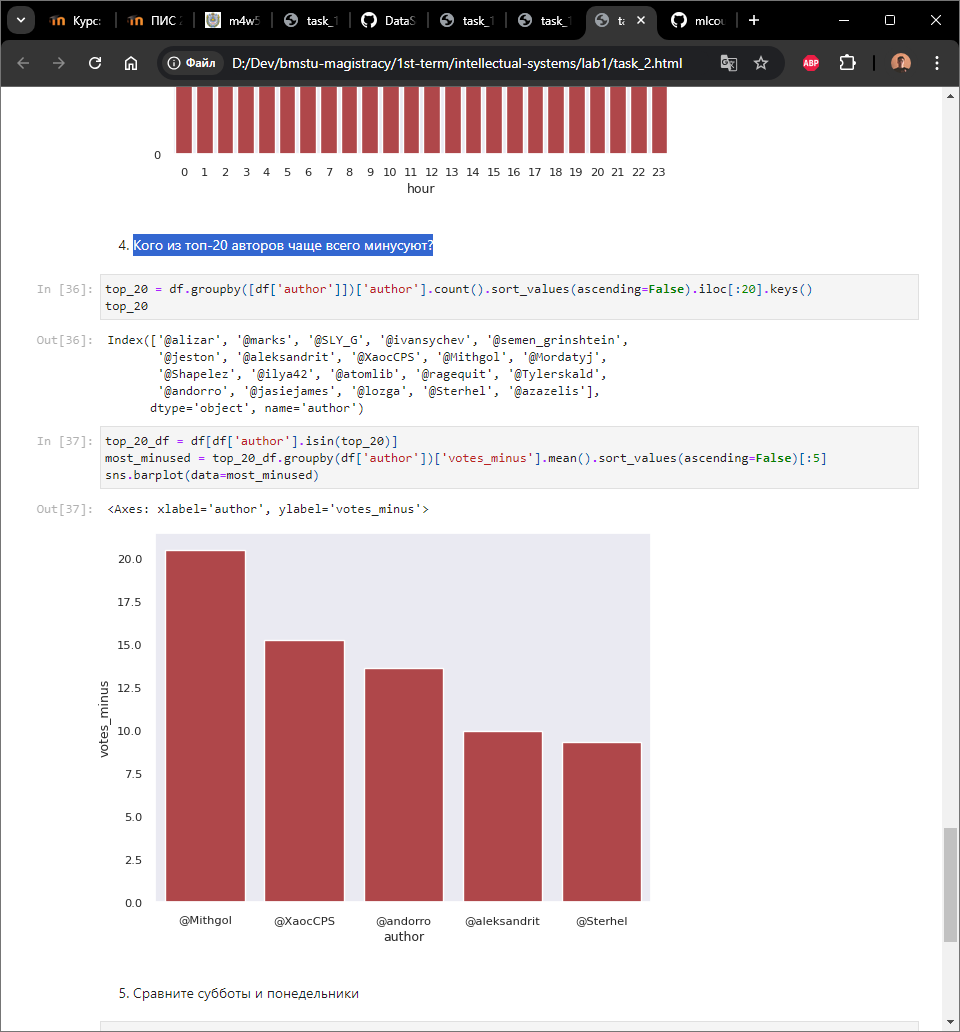
**Рисунок 23 –** Максимальные просмотры статей, опубликованных в заданный час, на гиктаймс

Таким образом, тезис «Максимальное число комментариев на гиктаймсе набрала статья, опубликованная в 9 часов вечера» опровергается.

Подтвердился лишь один тезис: «*Больше всего просмотров набирают статьи, опубликованные в 6 часов утра*».

**Задача 4**

Кого из топ-20 авторов (по количеству статей) чаще всего минусуют?

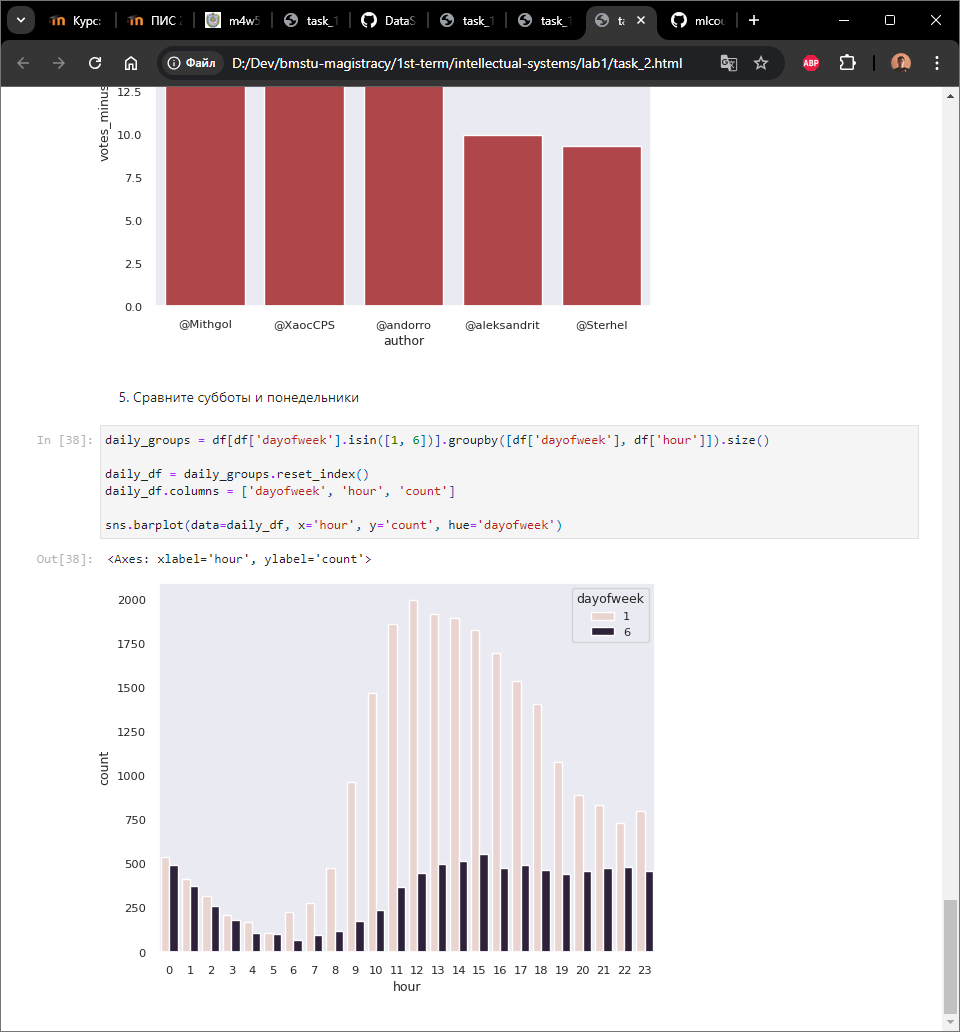


**Рисунок 24 –** Среднее количество минусов на статью для авторов

Таким образом, самым минусуемым автором из топ-20 является «Mithgol».

**Задача 5**

Сравните субботы и понедельники. Правда ли, что по субботам авторы пишут в основном днём, а по понедельникам — в основном вечером?



**Рисунок 25 –** Количество публикаций по дням недели для суббот и понедельников

Таким образом, по понедельникам и субботам пишут приблизительно в одно и то же время, но разные количества статей, так что тезис опровержен.

**Вывод**: в ходе выполнения лабораторной работы были получены практические навыки работы с большими массивами данных в pandas, а также навыки их визуализации с использованием библиотеки seaborn.