|  |  |
| --- | --- |
|  | **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, искусственный интеллект и системы управления»

КАФЕДРА «Системы обработки информации и управления»

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***Система оперативного анализа данных***

***«Приют для животных в Блумингтоне»***

Студент ИУ5-55Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Акулова А.А.

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

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

(Подпись, дата) (И.О.Фамилия)

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

(Подпись, дата) (И.О.Фамилия)

*2022 г.*

**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)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

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

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_\_\_\_\_\_\_Анализ данных по базе данных «Приют для животных в Блумингтоне»\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_\_\_\_\_ИУ5-55Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Акулова Анастасия Алексеевна\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Производственная\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к 3 нед., 50% к 9 нед., 75% к 12 нед., 100% к 15 нед.

***Техническое задание***  *Спроектировать систему анализа и визуализации по теме «Приют для животных в Блумингтоне», провести визуализацию полученных данных посредством языка программирования Python*

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 20 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

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

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

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

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

Дата выдачи задания «10» сентября 2022 г.

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

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** А.А. Акулова

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**ОГЛАВЛЕНИЕ**

[ВВЕДЕНИЕ 5](#_Toc122768087)

[1 Определение данных для анализа 6](#_Toc122768088)

[2 Формулирование гипотез 7](#_Toc122768089)

[3 Общее исследование данных 8](#_Toc122768090)

[3.1 Получение информации о данных 10](#_Toc122768091)

[3.2 Исследовательский анализ 12](#_Toc122768092)

[4 Анализ результатов 14](#_Toc122768093)

[5 Анализ DataFrame по умершим животным 17](#_Toc122768094)

[6 Какие животные чаще всего умирают 19](#_Toc122768095)

[7 Гипотеза. Чаще всего умирают бродячие кошки 21](#_Toc122768096)

[8 Гипотеза: Чаще всего собаки умирают от бешенства 25](#_Toc122768097)

[9 Живые животные 30](#_Toc122768098)

[10 Гипотеза: В основном в приют попадают взрослые и молодые животные 32](#_Toc122768099)

[11 Общий вывод 34](#_Toc122768100)

[ЗАКЛЮЧЕНИЕ 35](#_Toc122768101)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 36](#_Toc122768102)

# ВВЕДЕНИЕ

Цель:

Исследовать базу данных и выявить закономерности в приюте, чтобы предположить результаты лучшего ухода за животными.

Задачи:

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 «Приют для животных в Блумингтоне».

В наборе данных представлены суммарное количество золотых, серебряных и бронзовых медалей за всю историю Олимпийских игр для каждой страны.

Данные содержат следующие столбцы:

1. Id - идентификатор
2. Intakedate – дата приема
3. Intakereason – причина приема
4. Istransfer – является переводом
5. Sheltercode – код приюта
6. Identichipnumber – идентификационный номер
7. Animalname – кличка животного
8. Breedname – название породы
9. Basecolour – базовый цвет
10. Speciesname – название вида
11. Animalage – возраст животного
12. Sexname – название пола
13. Location - местоположение
14. Movementdate – дата перемещения
15. Movementtype – тип перемещения
16. Istrial – является испытательным сроком
17. Returndate – дата возврата
18. Returnedreason – причина возврата
19. Deceaseddate – дата смерти
20. Deceasedreason - умерший
21. Diedoffshelter – умер в приюте
22. puttosleep - усыплен

# Формулирование гипотез

Гипотеза 1:

*Нулевая гипотеза H0*: Чаще всего умирают бродячие кошки

*Альтернативная гипотеза H1:* Чаще всего умирают домашние кошки

Гипотеза 2:

*Нулевая гипотеза:* Чаще всего собаки умирают от бешенства

*Альтернативна гипотеза:* Чаще всего собак усыпляют

Гипотеза 3:

*Нулевая гипотеза:* В основном в приют попадают взрослые и молодые животные

*Альтернативна гипотеза:* Чаще всего в приют попадают животные средних лет

# Общее исследование данных

Анализ будет проводиться с помощью средств Python, используются библиотеки: pandas, numpy, matplotlib, seaborn, scipy, warnings, plotly, math. Python – наиболее удобная и быстрая среда для анализа данных. Удостоверимся, что все библиотеки загружены и подключены. Загрузим данные из csv файла в DataFrame.

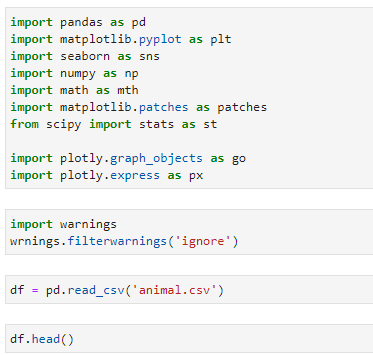


Рисунок 1 – Загрузка данных

Удостоверимся, что данные загружены нормально и посмотрим на получившуюся таблицу, проанализируем первые 5 строк, удостоверимся, что числа загрузились нормально – максимальные значения достигнуты и не пропущены.



Рисунок 2 - Изображение таблицы

# **3.1 Получение информации о данных**

Проверим наличие пропусков в данных:

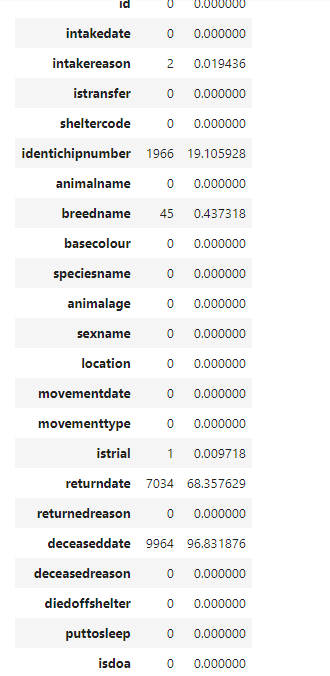


Рисунок 3 - Проверка данных

Пропуски выявлены в столбцах идентификационный номер животного, порода, испытательный срок, дата возврата, дата смерти.

* Пропуски в дате смерти может означать, то, что на данный момент сбора информации животное было живо, что не может не радовать, ведь их количество подавляющее большинство
* Пропуски в породе - может означать, что это бездомное животное - без породистое
* Пропуски в дате возврата - означают, что на данный момент сбора информации животное еще находилось в приюте
* Пропуски в id - не информативны, поэтому данный столбец можно не брать во внимание

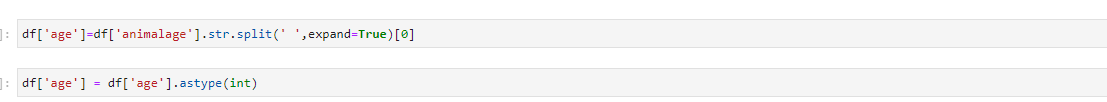


Рисунок 4 – Замена пропусков

# 3.2 Исследовательский анализ

Сделаем дополнительный DataFrame, в котором будем хранить информацию об умерших собаках, чтобы в дальнейшем проанализировать и выявить закономерность их смерти.

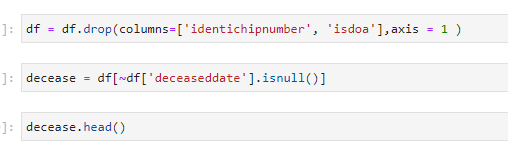


Рисунок 5 – DataFrame об умерших собаках

Удостоверимся, что данные загружены нормально и посмотрим на получившуюся таблицу, проанализируем первые 5 строк, удостоверимся, что числа загрузились нормально – максимальные значения достигнуты и не пропущены.

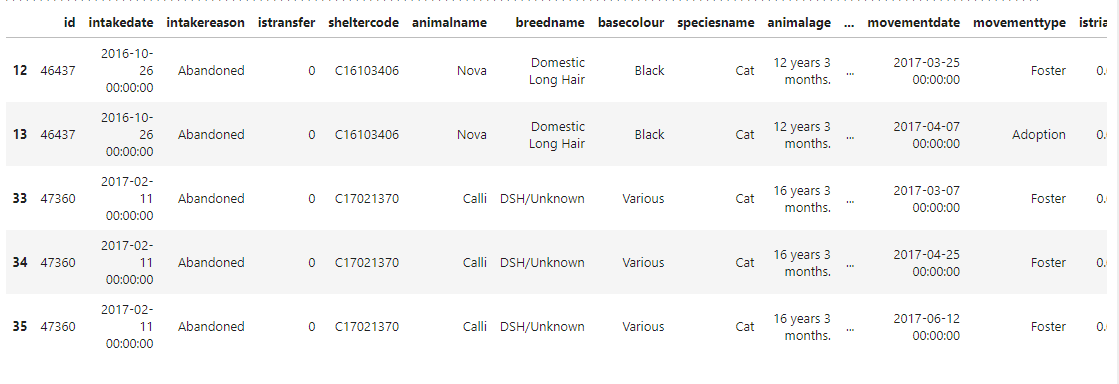


Рисунок 6 - Изображение таблицы

Сделаем дополнительный DataFrame, в котором будем хранить информацию об оставленных собаках, чтобы в дальнейшем проанализировать и выявить закономерность, почему их не забирают.

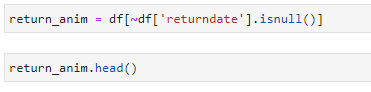


Рисунок 7 – DataFrame об оставленных собаках

Удостоверимся, что данные загружены нормально и посмотрим на получившуюся таблицу, проанализируем первые 5 строк, удостоверимся, что числа загрузились нормально – максимальные значения достигнуты и не пропущены.

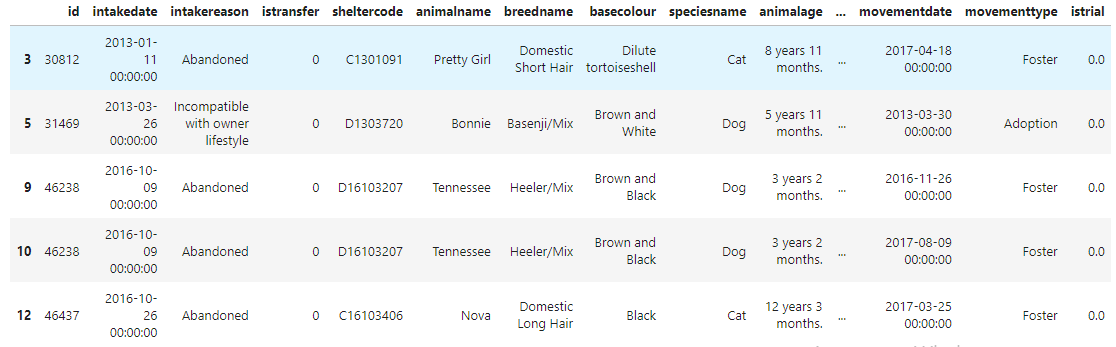


Рисунок 8 - Изображение таблицы

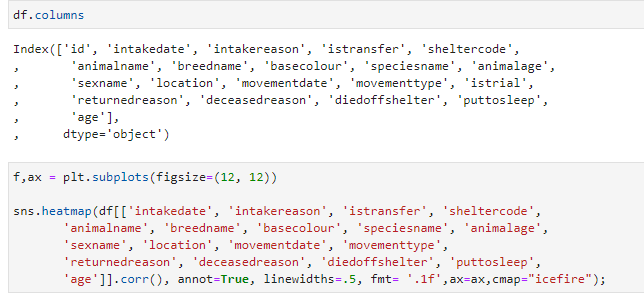
Удалим из DataFrame столбцы про смерть собаки и была ли она отдана новым хозяевам. Отдельно проанализируем эти столбцы

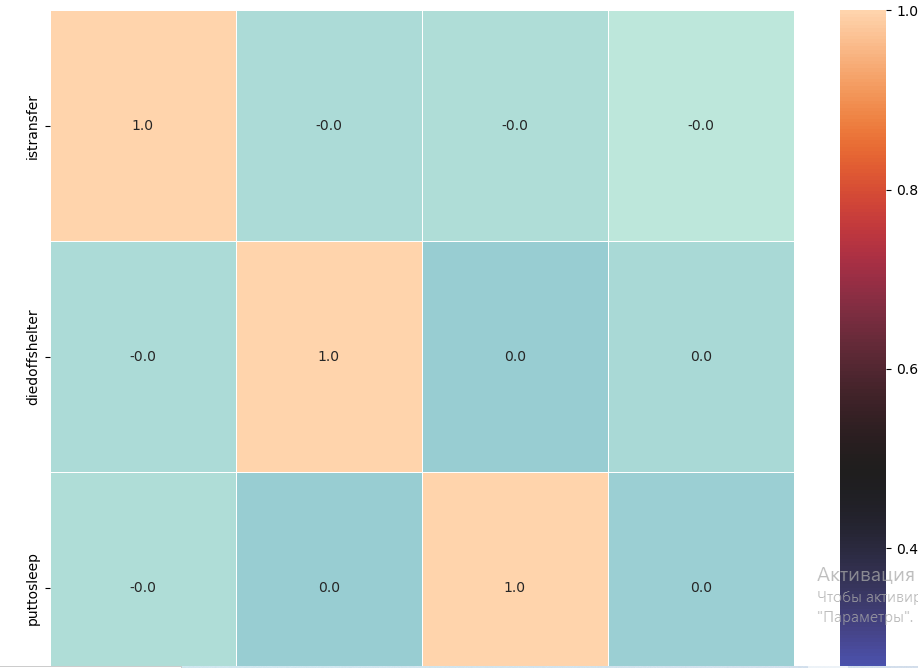


# 4 Анализ результатов

Анализ диаграммы, которая описывает тепловые связи.

Создадим DataFrame и изобразим получившийся результат на графике:





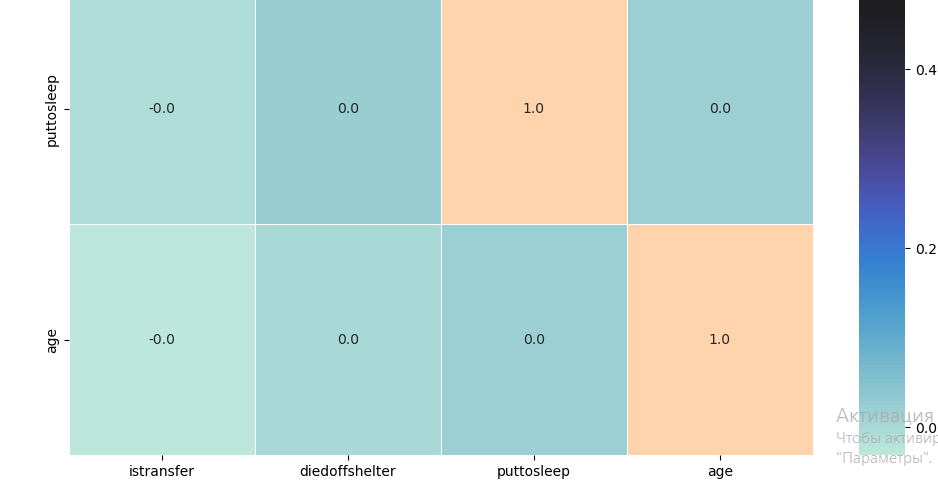


Рисунок 9 – Корреляция

Создадим график временной зависимости



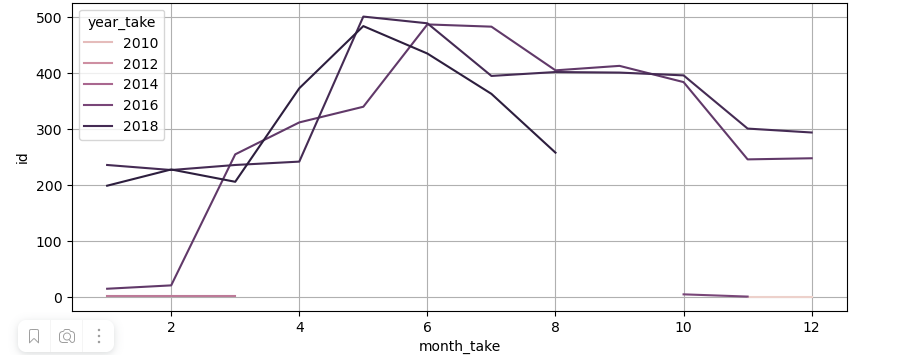


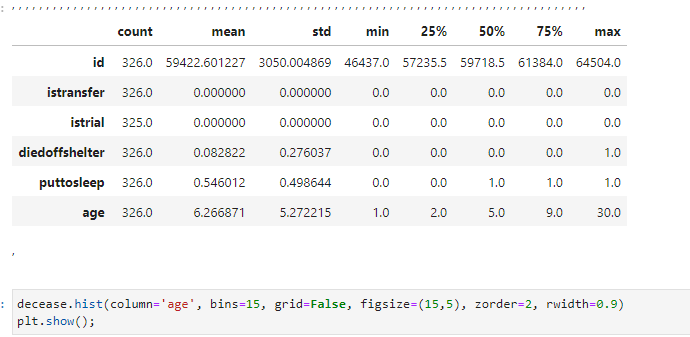
Рисунок 10 – График времени

Похоже, что каждый год весной в приюте будет больше приемных животных, вероятно, потому, что в это время животные рожают, что приведет к увеличению их числа.

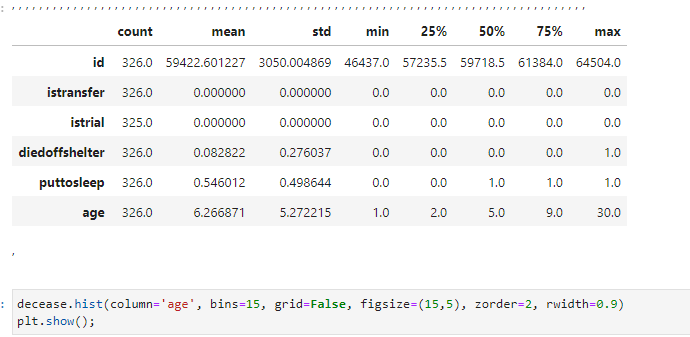
# 5 Анализ DataFrame по умершим животным

Выведем числовые данные





Построим диаграмму по числовым данным



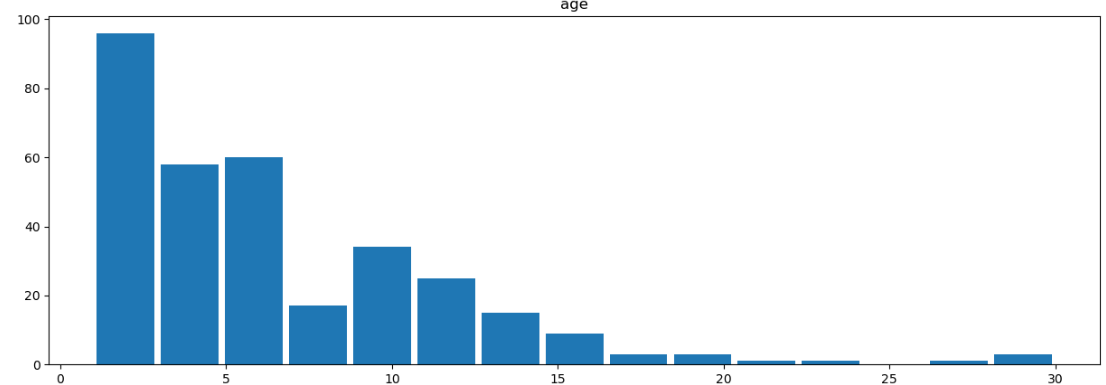
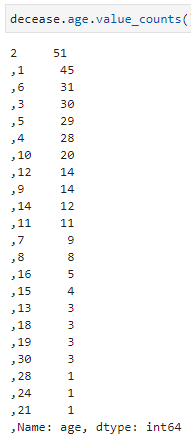


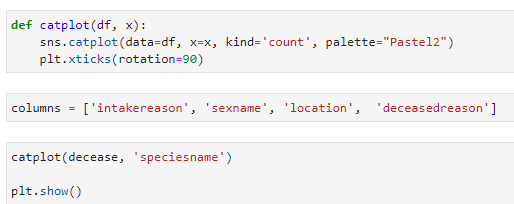
Рисунок 11 – Статистика по умершим животным



Грустная статистка, но факт остается фактом, в основном умирают животные возраста от **2 до 6 лет**

# 6 Какие животные чаще всего умирают

Определим у кого из животных наибольшая смертность.



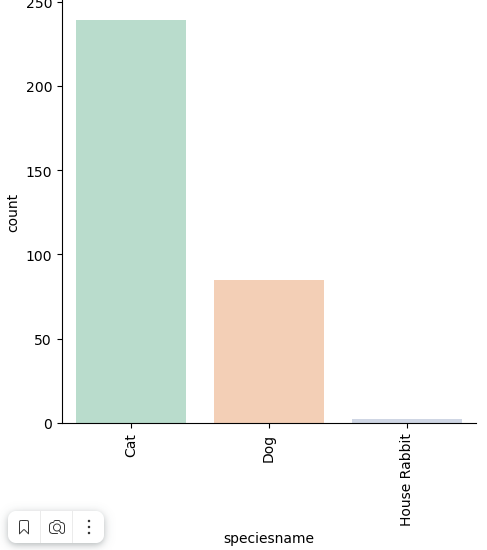
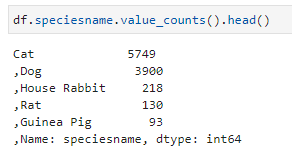


Рисунок 12 – Смертность животных

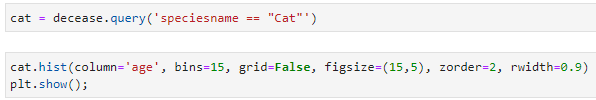


Так как кроликов в приюте подавляющее меньшинство, то не будем анализировать их смертность.

Сделаем срез данных по кошкам и собакам

# 7 Гипотеза. Чаще всего умирают бродячие кошки

Сделаем срез по кошкам.



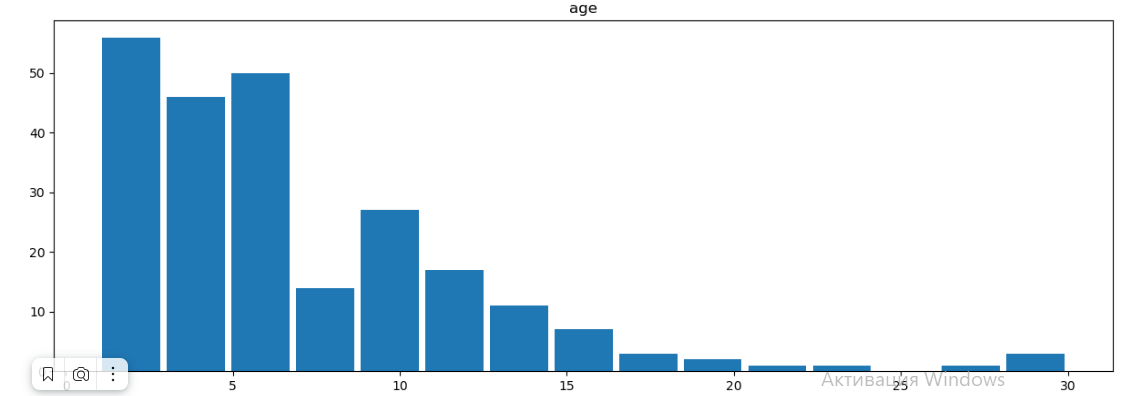
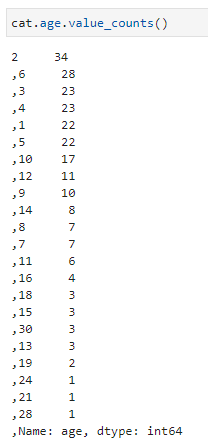
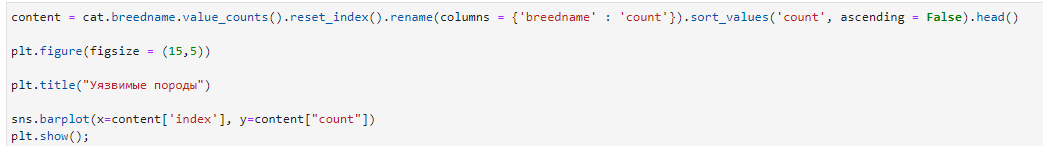


Рисунок 13 – Смертность кошек



Кошки чаще всего умирают в возрасте от **2 до 6 лет**

**Выявим уязвимые породы**



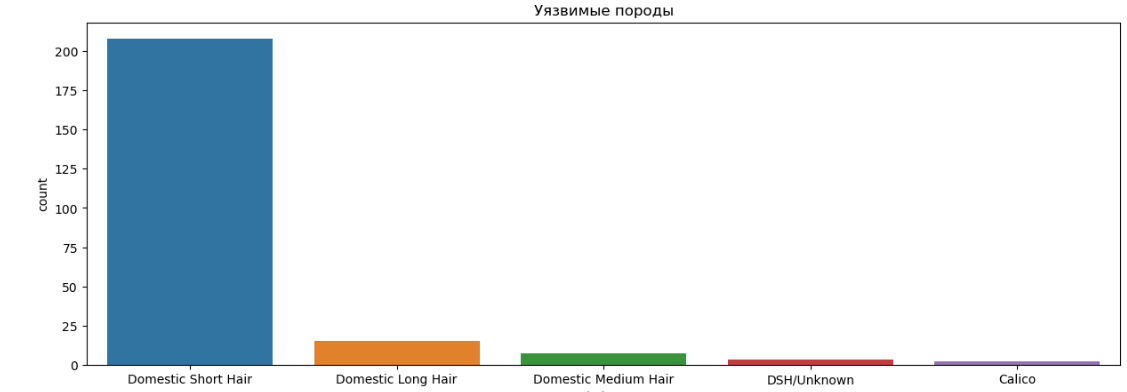


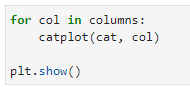
Рисунок 14 – Уязвимые породы

Среди кошек чаще всего умирают домашние короткошерстные животные.

Чаще всего умирают бродячие кошки, мужского пола, в вальере номер А, которые были возвращены в приют, но при этом умирали они в заботе

Это легко объясняется тем, что бродячие кошки более уязвимы к лишаю и прочим заболеваниям, которым должны лечиться в ветклиниках, чего нет в приютах

Гипотеза подтвердилась.



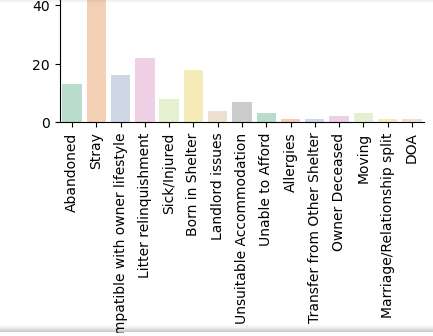


Рисунок 15 – Причина попадания в приют у кошек

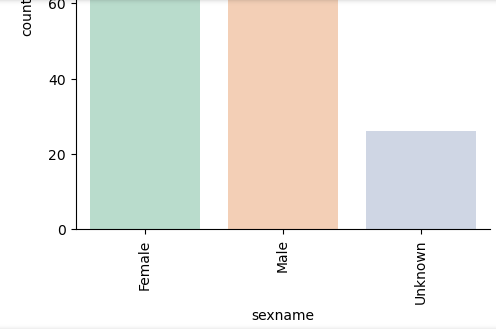


Рисунок 16 – Пол попавших в приют кошек

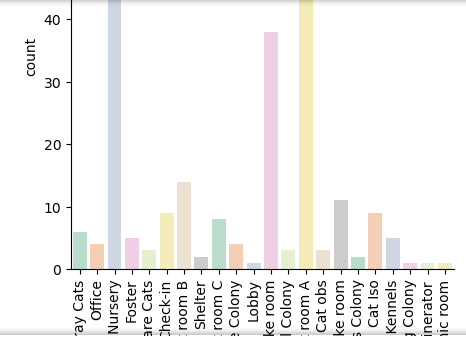


Рисунок 17 – Местонахождения кошек

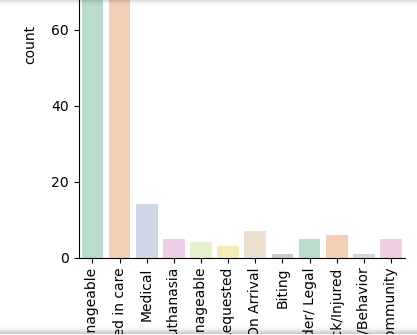
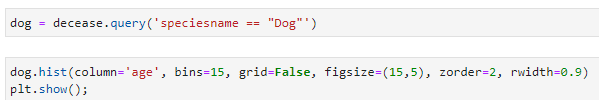


Рисунок 18 – Причины смерти у кошек

# Гипотеза: Чаще всего собаки умирают от бешенства

Сделаем срез по собакам.



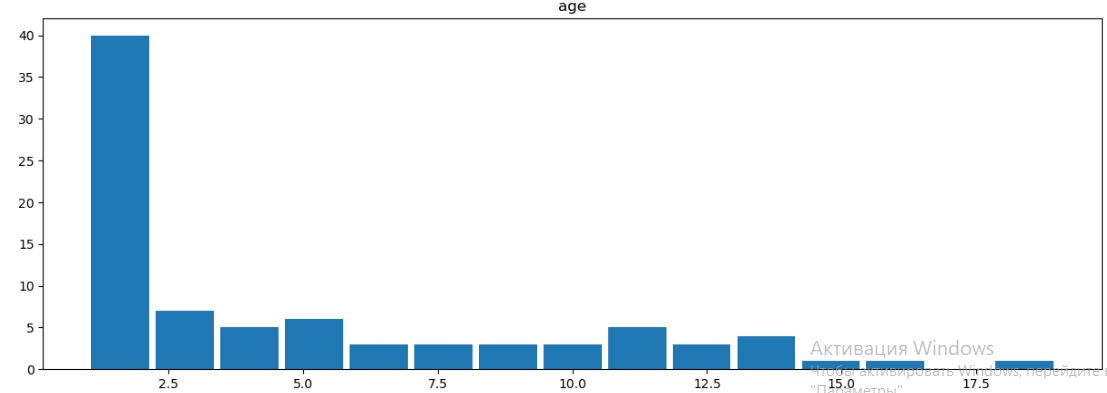
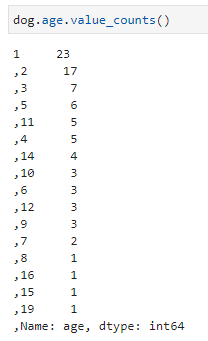


Рисунок 19 – Смертность собак



Статистика по собакам намного хуже, чем у кошек чаще всего они умирают в возрасте **от 1 до 2 лет**

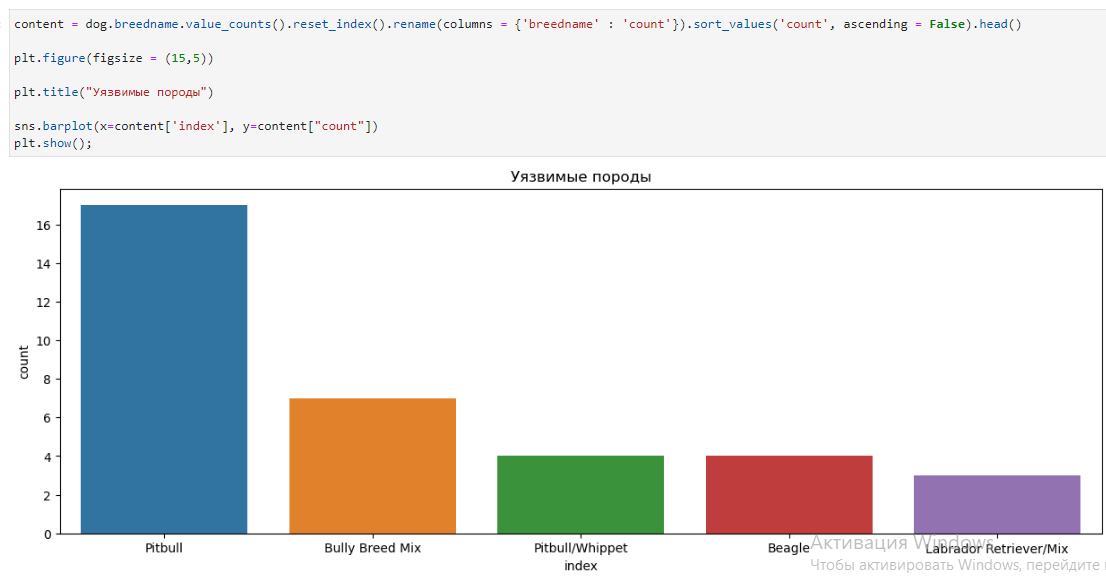


Рисунок 20 – Уязвимые породы собак

Среди собак чаще всего умирают питбули.

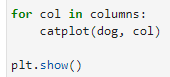
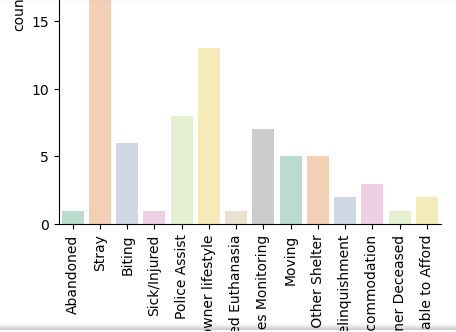
 

Рисунок 21 – Причины попадания собак в приют

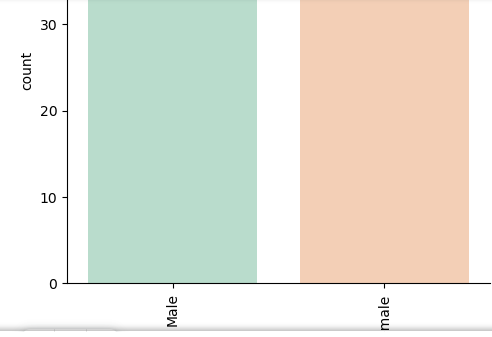


Рисунок 22 – Пол собаки в приюте

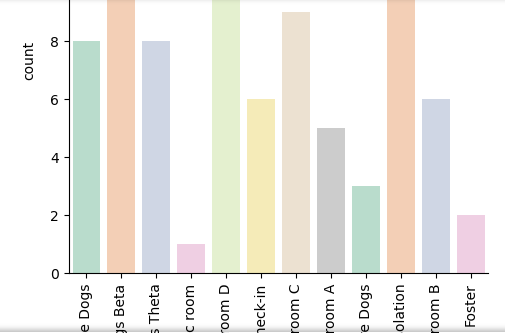


Рисунок 23 – Местонахождения собаки

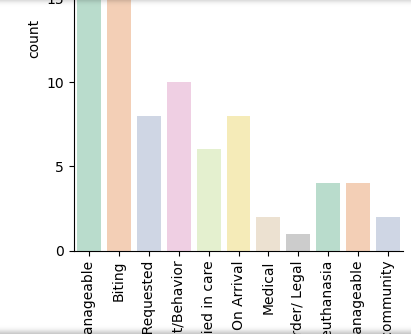


Рисунок 24 – Причина смерти собаки

Чаще всего умирают бродячие собаки, мужского пола, в вальере номер D из-за усыпления от бешенства.

Заразиться бешенством можно лишь через кровь, слюну, биологические жидкости. Нельзя его подхватить через мочу, съев или понюхав что-то. Так собаки более агресивны и чаще впадают в драки между друг другом вероятность заразиться у них выше, поэтому и смерность по этому поводу у них выше

Гипотеза подтвердилась.

Вывод

1. Чаще всего в приюте умирают кошки, но это скорее всего можно объяснить тем, что в самой выборке датафрейма кошек в 1,5 раза больше, чем собак, поэтому считать, что это закономерность из-за типа животных не стоит
2. Подавляющее большинство умерло бродячих животных
3. И если причина смерти у кошки была смерть в заботе, те скорее всего естественным причинам, то собак чаще всего усыпляли из-за бешенства

# Живые животные

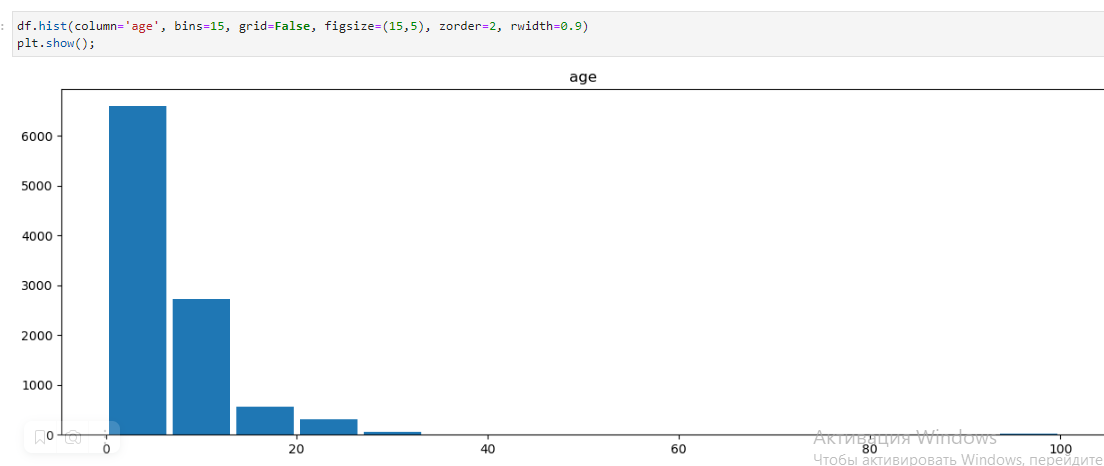
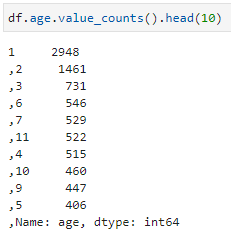


Рисунок 25 – Статистика по живым животным



Подавляющее большинство животных находятся в приюте в возрасте от 1 до 2 лет. Сделаем категоризацию данных по возрасту, что бы выдвинуть и проанализировать гипотезы связанные с возрастом животных.



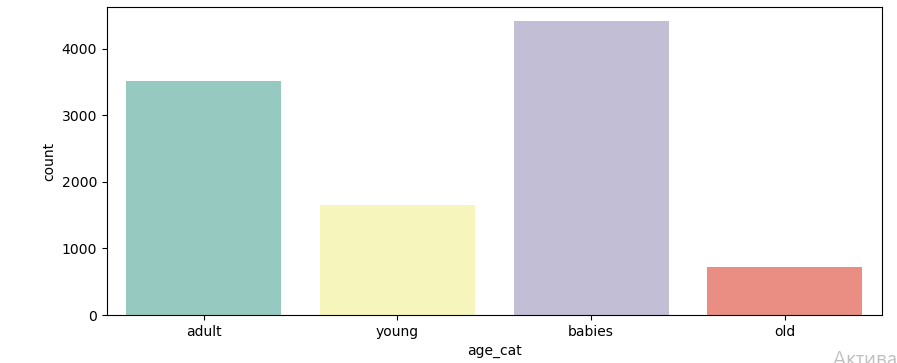
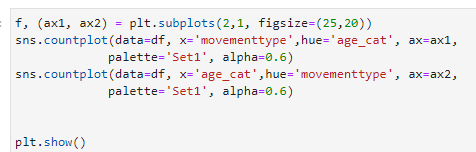


Рисунок 26 – Возраст животных



# Гипотеза: В основном в приют попадают взрослые и молодые животные



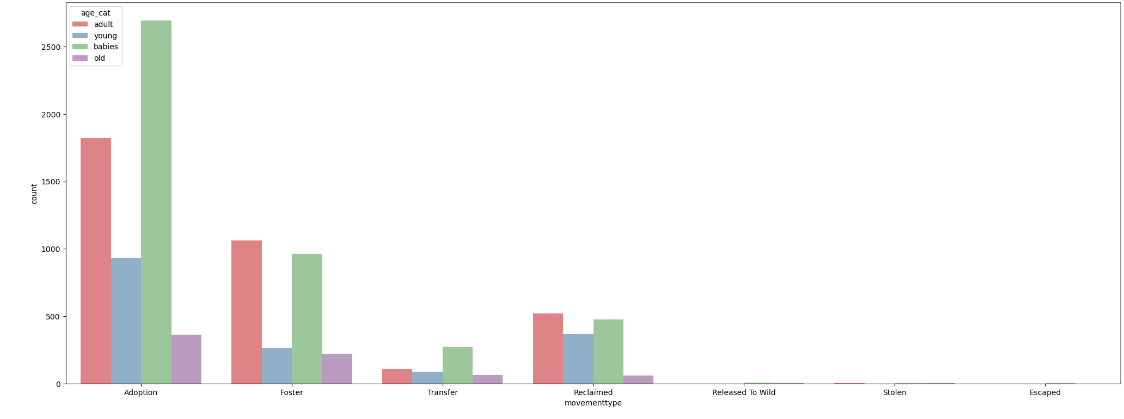


Рисунок 27 – Как животные попали в приют

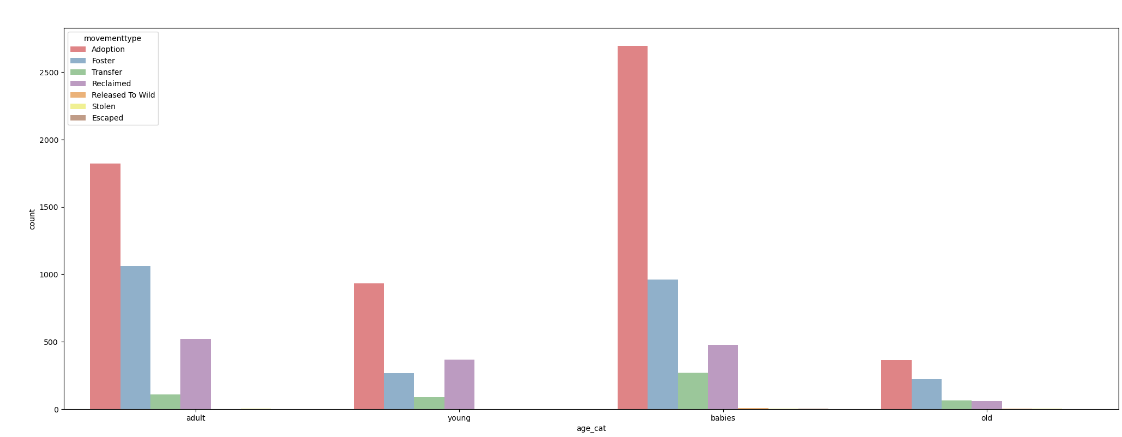


Рисунок 26 – Возраст животного, попавшего в приют

Как мы и думали чаще всего бросают молодых и старых животных. Это можно объяснить тем, что чаще всего люди, заводящие домашнее животное, не осознают всей ответственности: они не готовы к тому, что, например, молодое животное может портить мебель, а на больного старого животного надо тратить деньги из семейного бюджета на лечение.

# Общий вывод

Перед анализом мы провели проверку данных на пропуски, дубликаты. Также мы проверили названия столбцов и обнаружили в них ошибки. Удалили лишние символы из столбца с кодом страны.

В ходе анализа мы выяснили:

* + 1. Какие животных чаще сдают в приют;
    2. Кто из животных чаще умирает;
    3. Причины смертности животных;
    4. Место смерти животного;

Также мы сформулировали и проверили гипотезы, получилось, что:

1. Чаще всего умирают бродячие кошки, мужского пола, в вальере номер А, которые были возвращены в приют, но при этом умирали они в заботе
2. Чаще всего умирают бродячие собаки, мужского пола, в вальере D номер из-за усыпления от бешенства.
3. Чаще всего бросают молодых и старых животых

# ЗАКЛЮЧЕНИЕ

В ходе научно-исследовательской работы был проведен анализ базы данных «Приют для животных в Блумингтоне», выявлены важные статистические данные, которые могут помочь лучше понять жизнь бездомных животных.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas» и «Seaborn» на базе языка программирования Python.

# 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»