Лабораторная Работа №3  
Выполнил студент группы ИИм - 231   
Трофимова Валерия  
Вариант 10

Часть 1

В данной части работа выполняется с набором данных Fish.csv.

1. Вывести датасет и типы его переменных

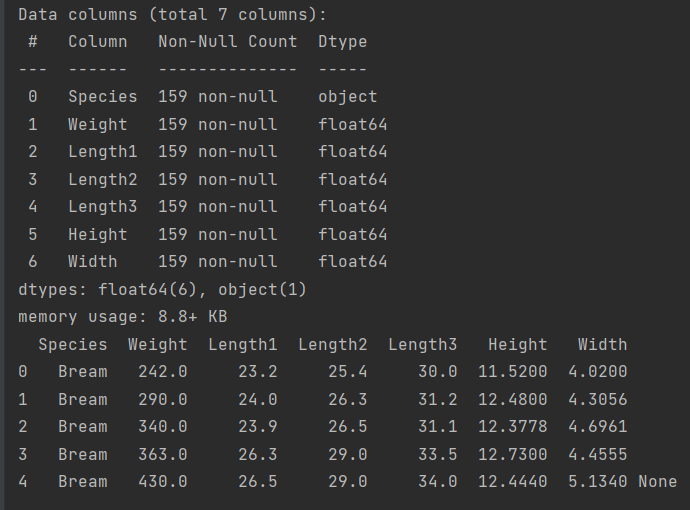


Рисунок 1 – Датасет и его переменные.

Как видно из рисунка выше, в датасете есть 7 переменных:

* Species – Вид рыбы.
* Weight – Вес рыбы.
* Length1 – Длина рыбы (Измерение в первом месте).
* Length2 – Длина рыбы (Измерение во втором месте).
* Length3 – Длина рыбы (Измерение в третьем месте).
* Height – Высота рыбы.
* Width – Ширина рыбы.

Все переменные имеют тип Float (числа с плавающей запятой), кроме переменной Species которая имеет тип - , или же в нашем случае String (строковый тип).

1. Используя визуализацию, покажите, сколько различных видов рыб включает ваш набор данных.

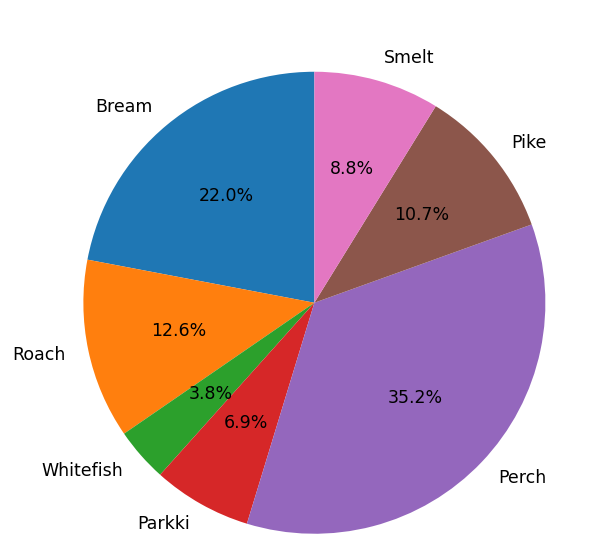


Рисунок 2 – Диаграмма распределения количества рыб по видам

Как видно из рисунка, всего в нашем датасете представлены 7 видов рыб:

* Perch – Окунь
* Bream – Лещ
* Roach – Плотва
* Pike – Щука
* Smelt – Корюшка
* Parkki – Ёрш
* Whitefish – Белая рыба

1. Исследуйте, есть ли какие - либо зависимые переменные.

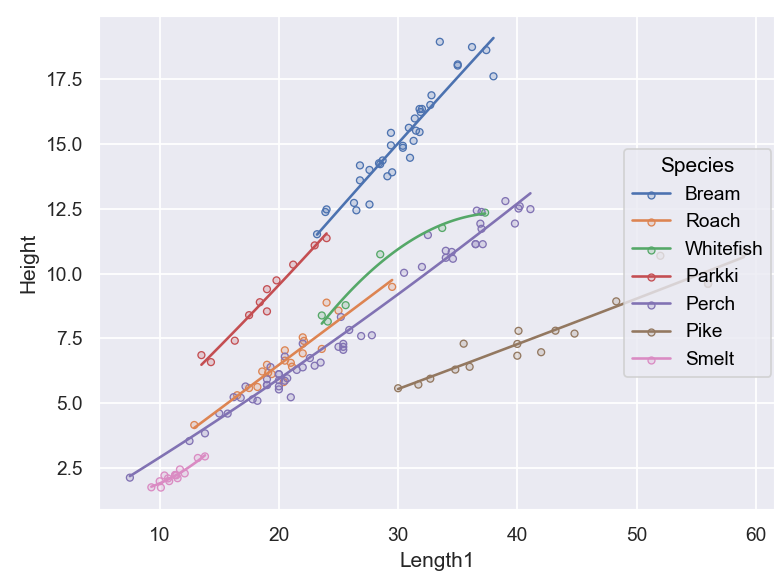


Рисунок 3 – Диаграмма рассеивания для параметров Length1 и Height.

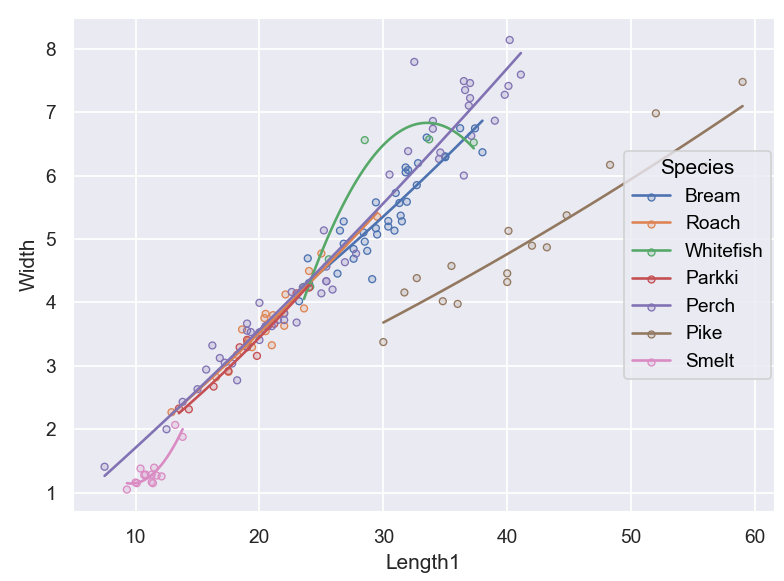


Рисунок 4 – Диаграмма рассеивания для параметров Length1 и Width.

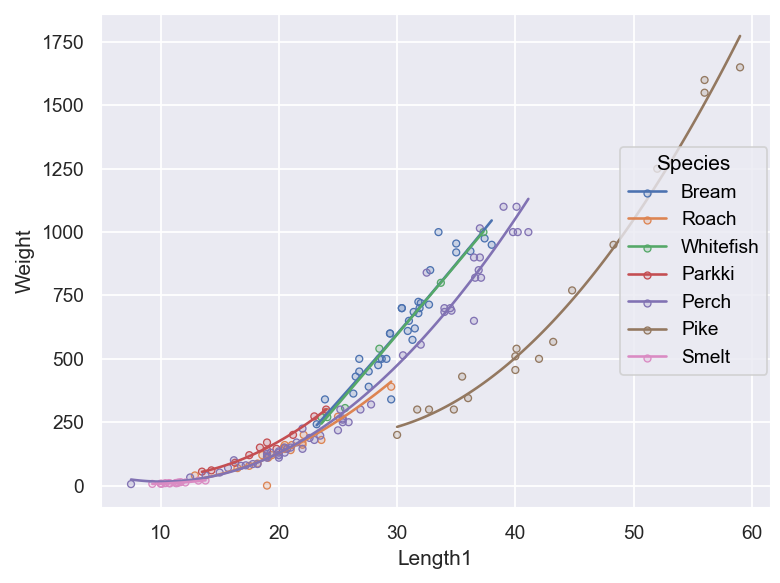


Рисунок 5 – Диаграмма рассеивания для параметров Length1 и Weight.

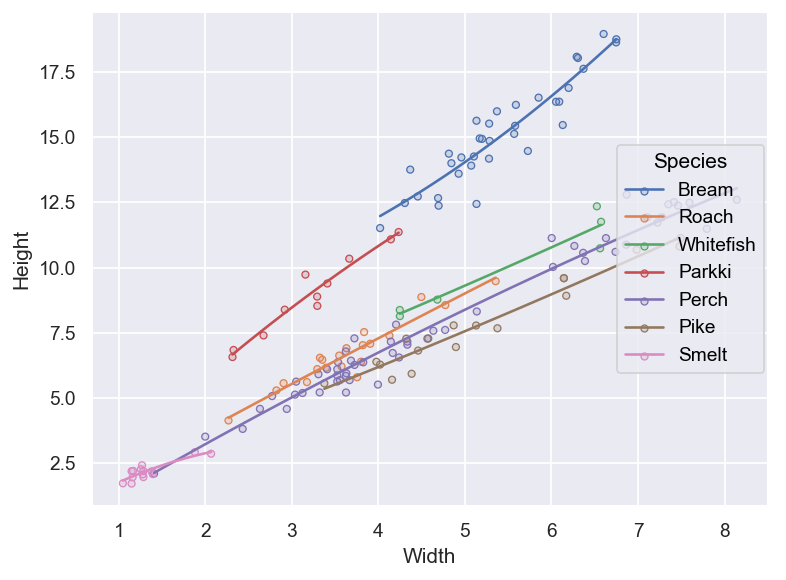


Рисунок 6 – Диаграмма рассеивания для параметров Width и Height.

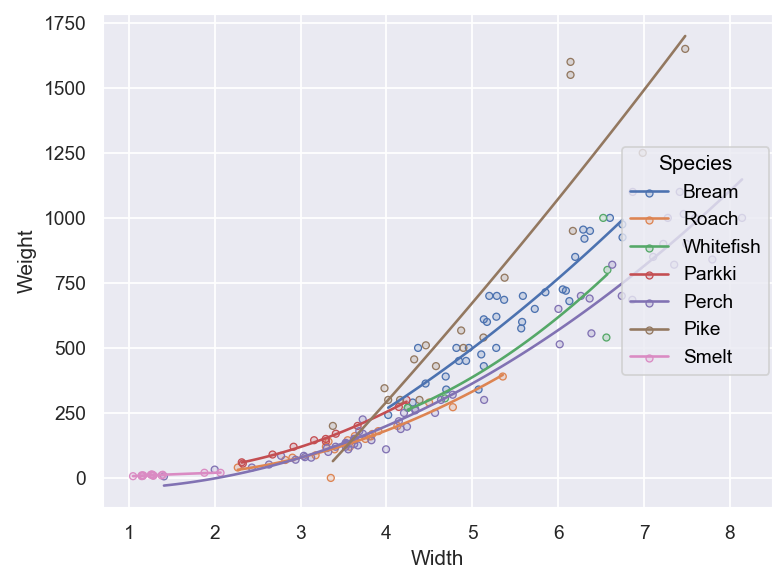


Рисунок 7 – Диаграмма рассеивания для параметров Width и Weight.

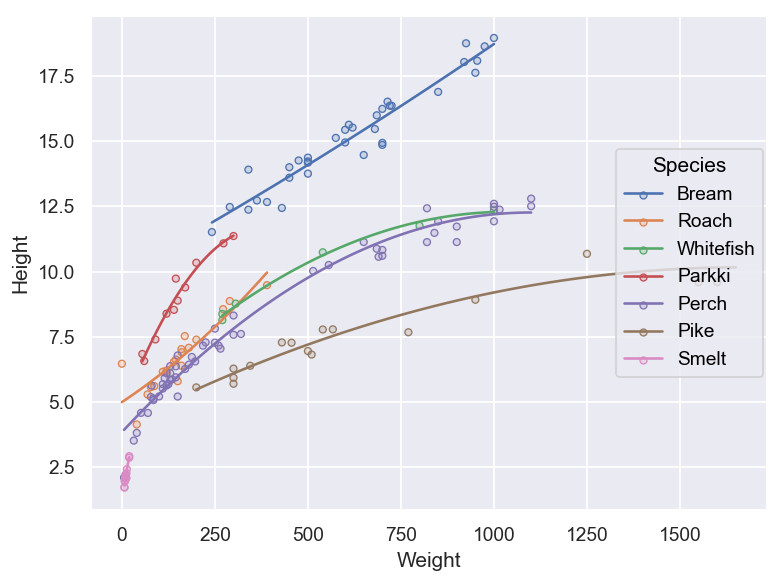


Рисунок 8 – Диаграмма рассеивания для параметров Weight и Height.

Нами были построены 5 диаграмм рассеивания для выявления зависимости между параметрами Length1, Weight, Height и Width. По построенным диаграммам можно сделать следующие выводы:

* Между Length1 и Height существует линейная зависимость, чем больше длинна, тем больше и высота рыбы;
* Между Width и Height существует линейная зависимость, чем больше ширина рыбы, тем больше её высота;
* Между параметрами Weight и Height также существует линейная зависимость, чем больше вес, тем больше высота;
* Ширина рыбы линейно зависима от длины рыбы;
* Длина рыбы нелинейно зависима от веса рыбы, при увеличении веса, значительно увеличивается длина;
* Ширина рыбы нелинейно зависима от веса рыбы, при увеличении веса, ширина значительно увеличивается.

1. Определите тип распределения количественных переменных

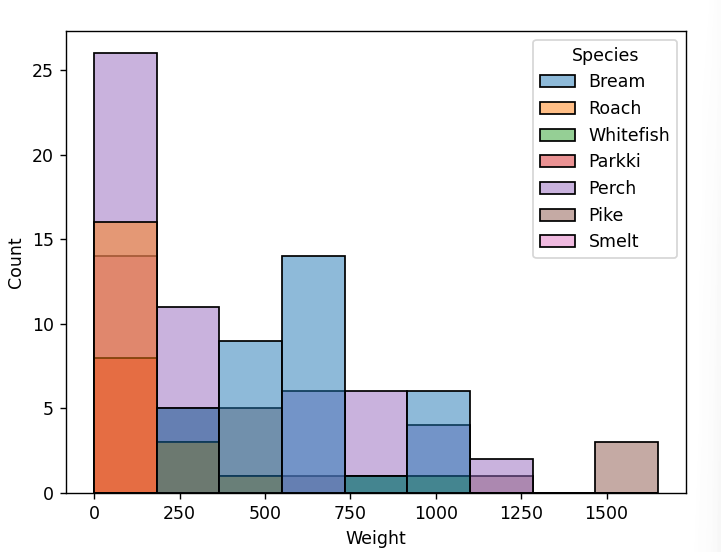


Рисунок 9.1 – Диаграмма распределения веса рыб.

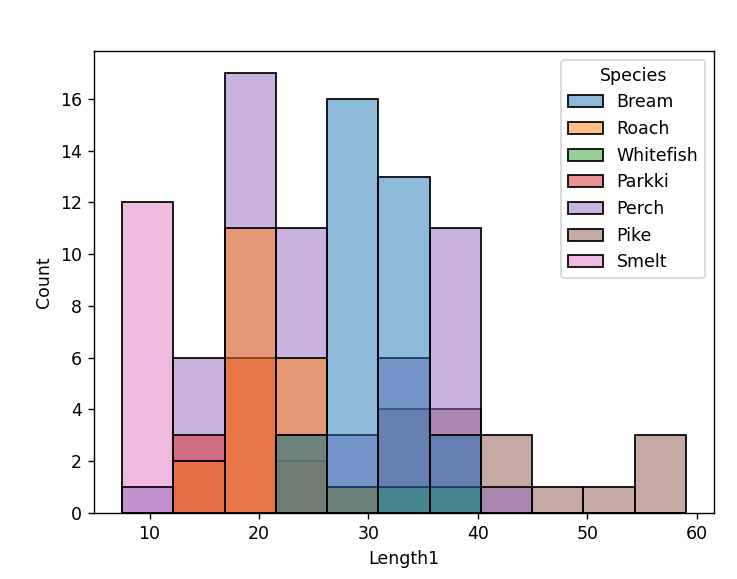


Рисунок 9.2 – Диаграмма распределения первой длинны рыб.

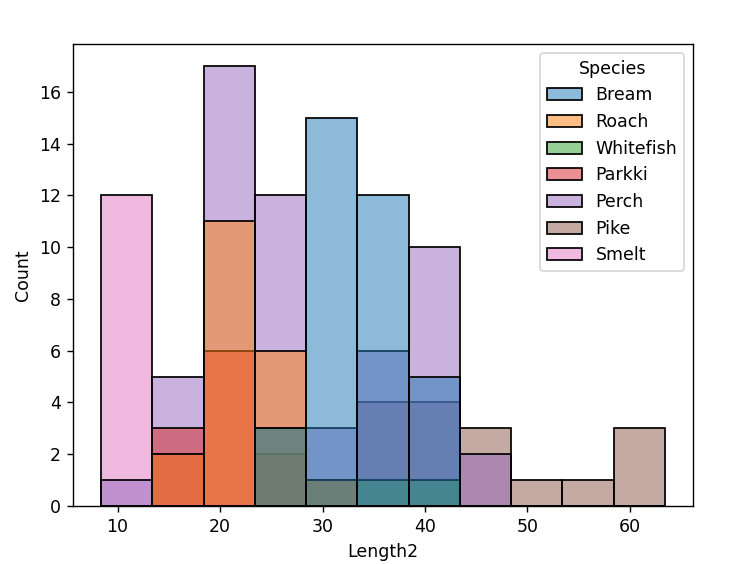


Рисунок 9.3 – Диаграмма распределения второй длинны рыб.

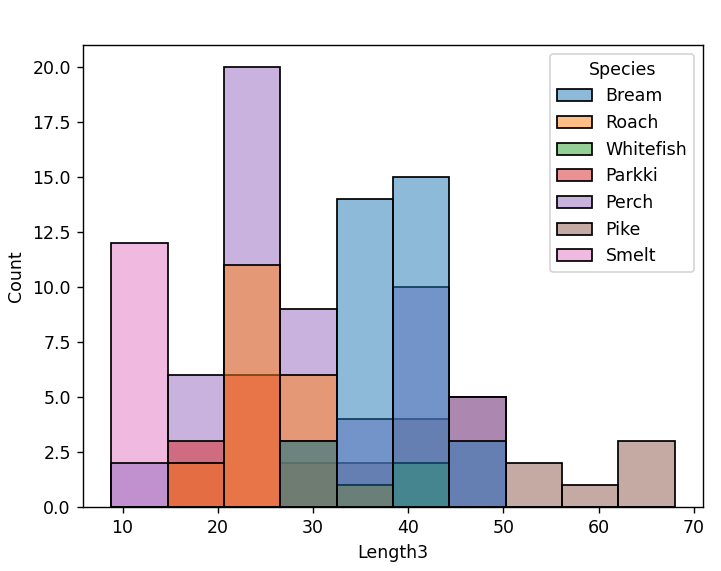


Рисунок 9.4 – Диаграмма распределения третьей длинны рыб.

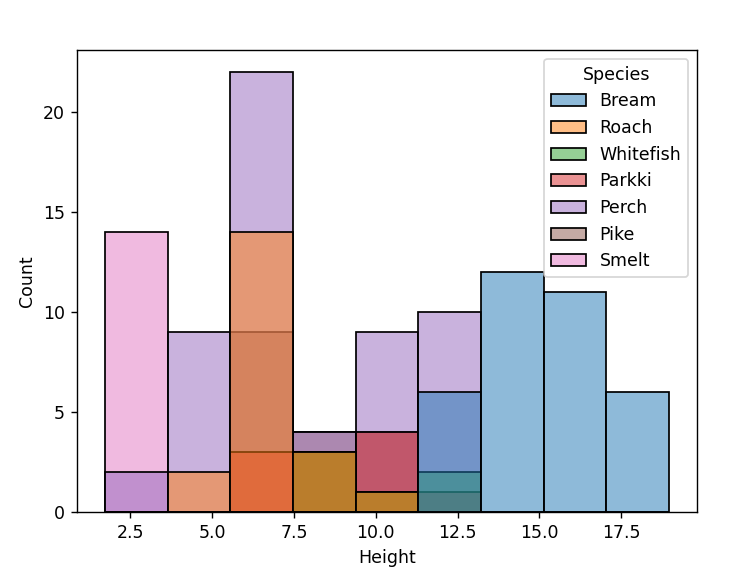


Рисунок 9.5 – Диаграмма распределения высоты рыб.

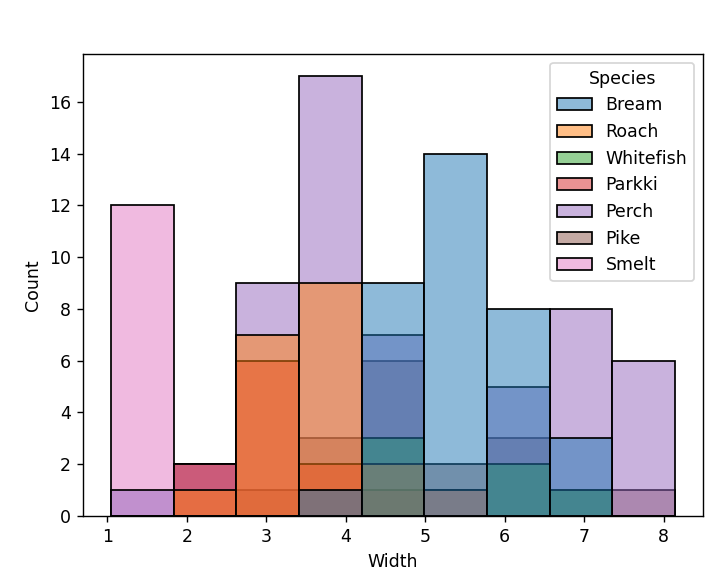


Рисунок 9.6 – Диаграмма распределения ширины рыб.

По полученным графикам можно сделать вывод что большая часть параметров имеет нормальное распределение, в то время как вес рыб имеет распределение Рэлея, при котором чем больше вес рыбы, тем меньше количество рыб с таким весом.

1. Существуют ли в этом наборе данных аномальные наблюдения?

Для определения наличия аномальных значений можно воспользоваться графиками из предыдущего пункта (Рисунки 9.1 – 9.6). На них можно увидеть распределение количества рыб, относительно значений определенного параметра. Ни на одном из графиков не присутствуют аномальные значений, сильно отличающиеся в большую или меньшую сторону от остальных значений, а значит можно сделать вывод об отсутствии аномальных значений.

1. Проверьте данные на наличие пропущенных значений.



Рисунок 10 – Определение наличия пропущенных значений.

Как видно из информационной таблицы, в наборе данных 159 строк, и у каждого из 7 параметров в нашем наборе присутствует 159 значений, а значит пропущенные значения в нашем датасете отсутствуют. Для наглядности построим график, на котором черным отображаются пропуски значений, а светло фиолетовым присутствующие значений параметров. График также подтверждает то, что в рассматриваемом наборе данных нет пропущенных значений.

1. Реализовать визуализацию многомерных данных.

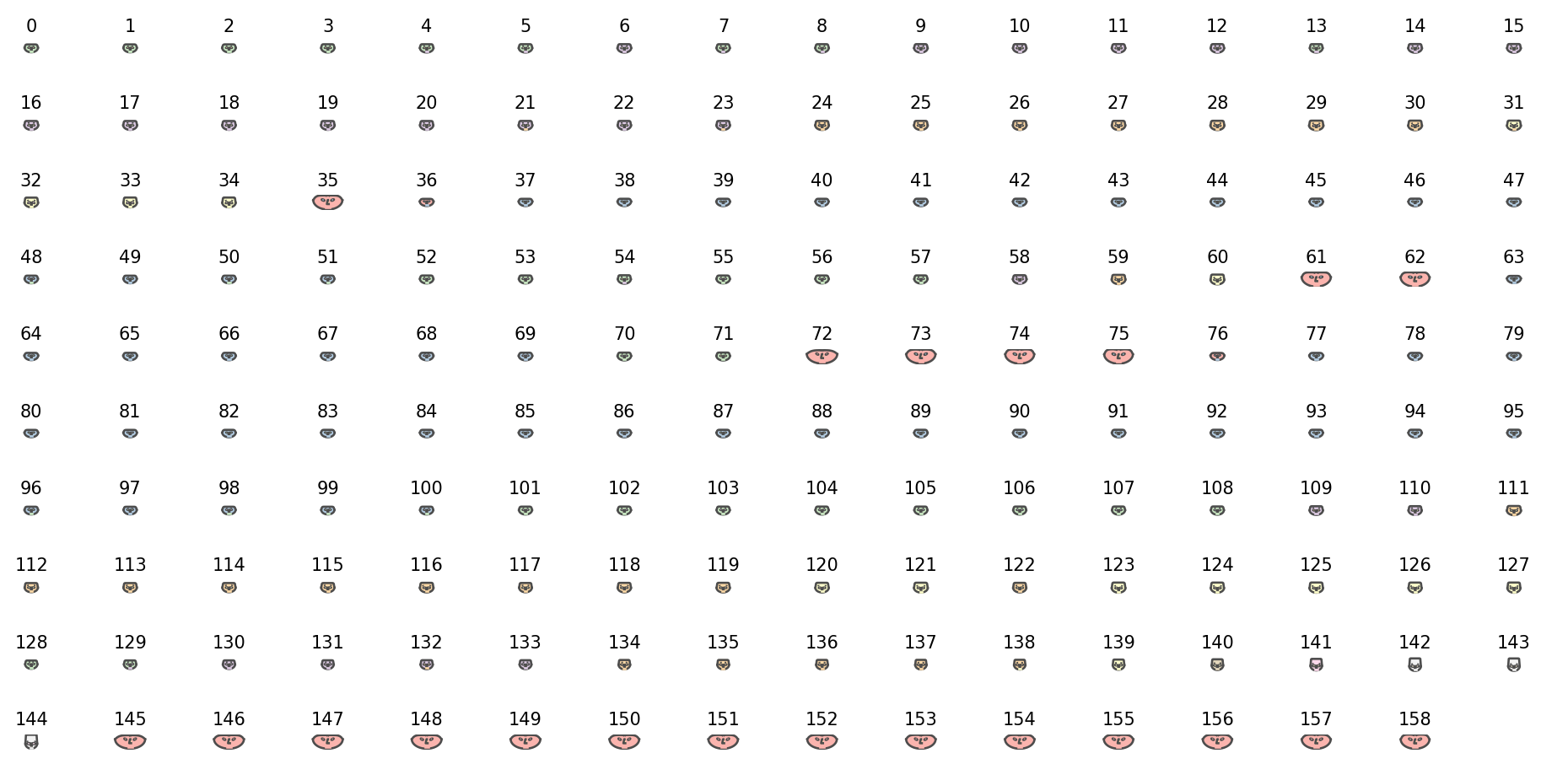


Рисунок 11.1 – Визуализация многомерных данных в виде лиц Чернова.

Как видно из Рисунка 11.1, график в виде лиц Чернова плохо подходит для визуализации наборов данных, состоящих из большого количества строк, так как для каждой строки строится своё лицо, и становится труднее оценивать каждый параметр, влияющий на внешний вид построенного лица. В данном случае можно лишь точно определить то, что на цвет лица влияет вид рыбы.

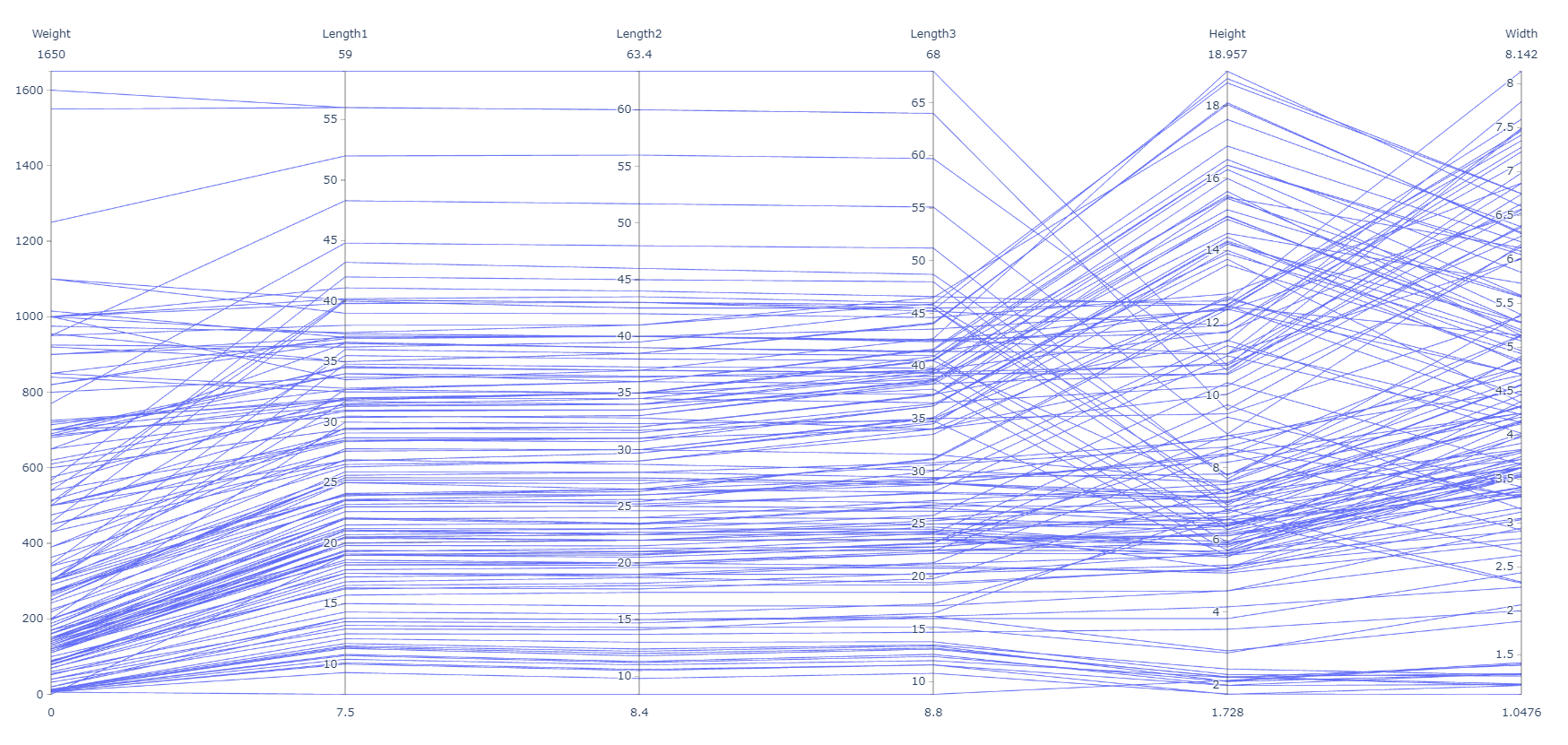


Рисунок 11.2 – Визуализация многомерных данных в виде параллельных координат.

Из данного рисунка можно сделать вывод, что все рыбы вида «Корюшка» обладают малыми размерами и весом, а самыми крупными являются рыбы вида «Щука».

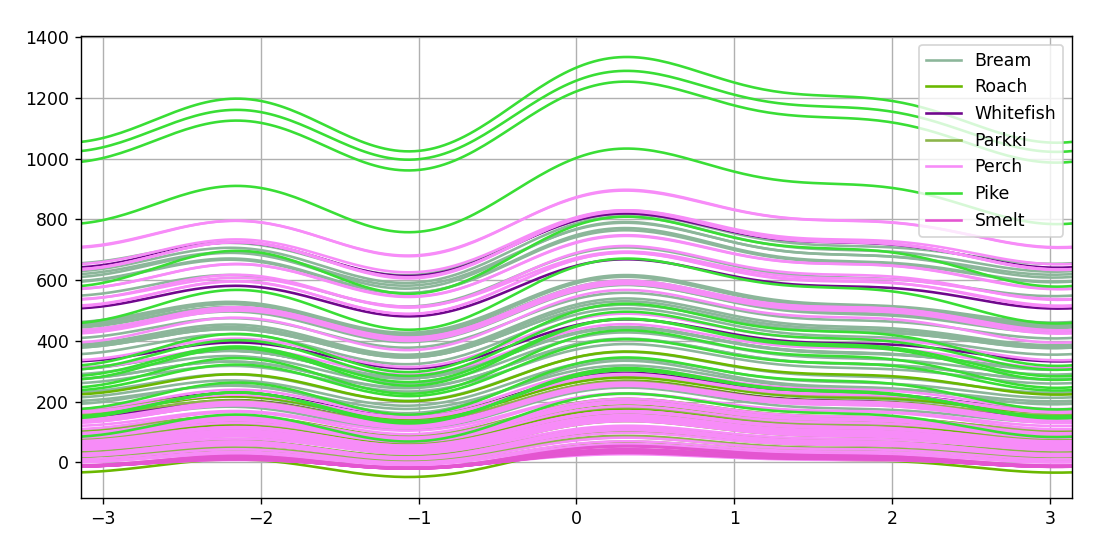


Рисунок 11.3 – Визуализация многомерных данных в виде кривых Эндрюса.

Как видно из данного графика, выводы по поводу наименьших и наибольших по размеру видов рыб, сделанных по предыдущему графику, подтверждаются.

1. Регрессионный анализ по 2 параметрам

Далее построим модель линейной регрессии, где целевой переменной (y) будет значение параметра Weight, а зависимой переменной (x) будет параметр Length1. Так как строить модель для всех видов рыб одновременно нецелесообразно из - за выявленных ранее различий в размерах различных видов рыб, то мы будем строить модель только для вида «Окунь» так как он является самым распространенным в нашем датасете.

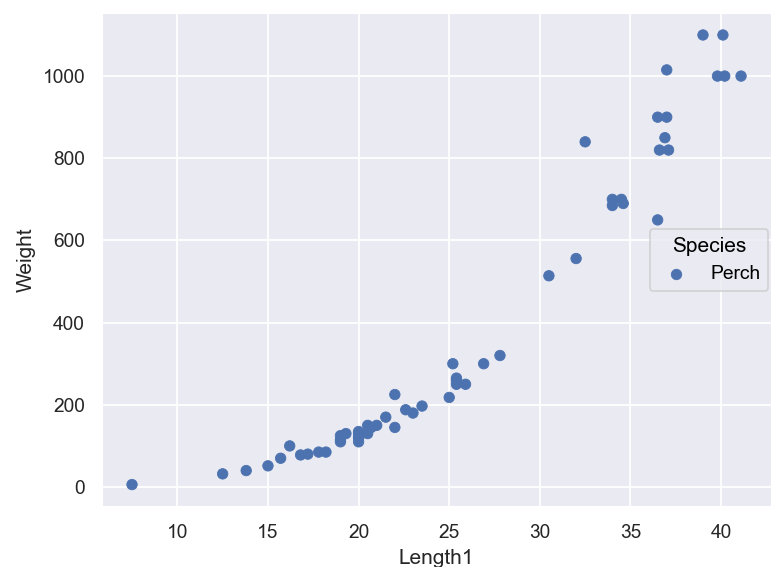


Рисунок 12 – Диаграмма рассеивания для всех пар значений.

Далее мы создаём модель линейной регрессии, разделив значения на тренировочные и тестовые части в пропорции 8 к 2. После, по полученной из тренировочных значений модели рассчитаем предсказанные значения для тестовой части значений, и сравним их с действительными значениями.

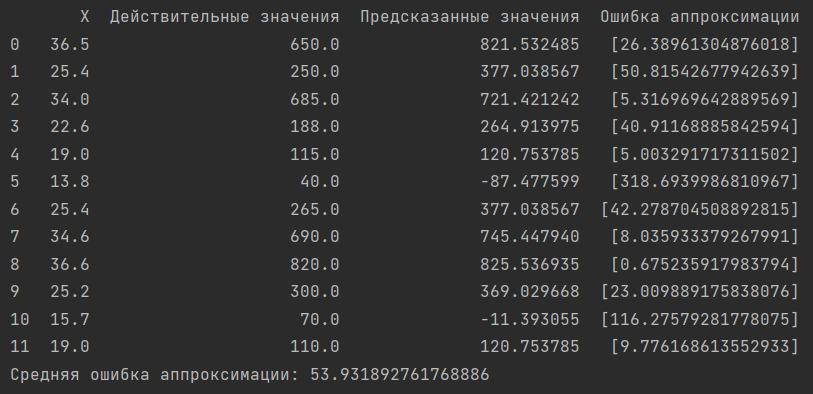


Рисунок 13 – Сравнение действительных и предсказанных значений.

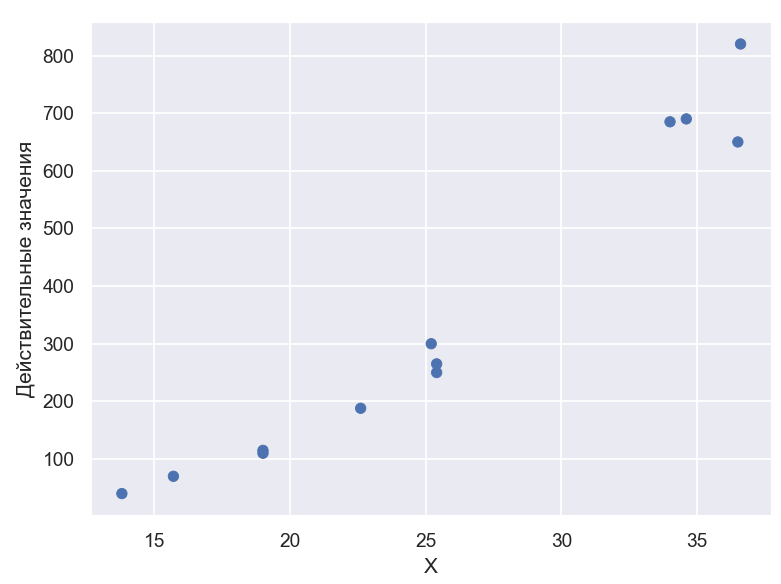
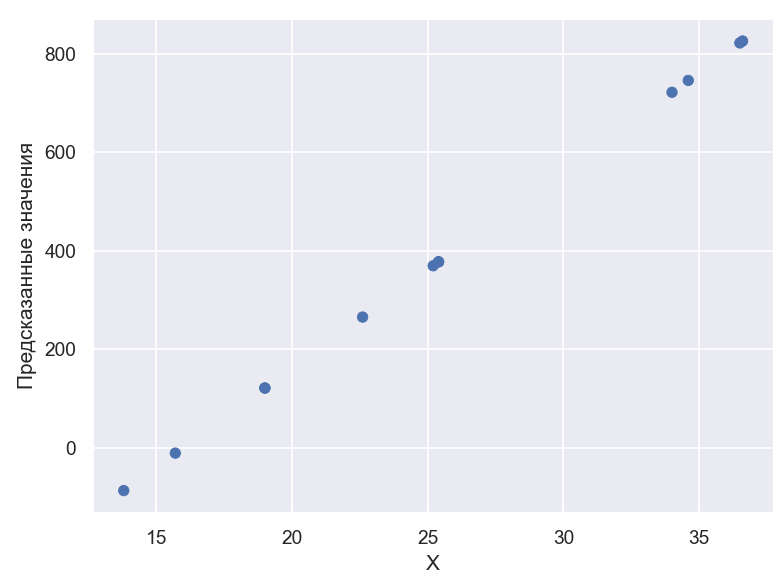
 

Рисунок 14 – Сравнение графиков действительных значений и предсказанных.

По полученным результатам можно увидеть, что средняя ошибка аппроксимации (отличие действительных значений от предсказанных) составляет 53,93%, что является плохим результатом. Наибольшие отличия наблюдаются при расчёте для малых значений длины рыбы.

1. Регрессионный анализ по нескольким параметрам

Повторим построение модели линейной регрессии из предыдущего пункта, только в этот раз в качестве зависимых переменных будут выступать все доступные атрибуты.

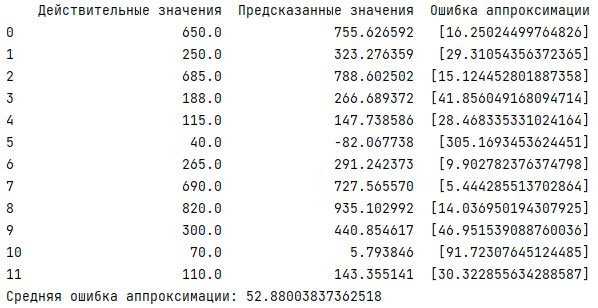


Рисунок 15 - Сравнение действительных и предсказанных значений.

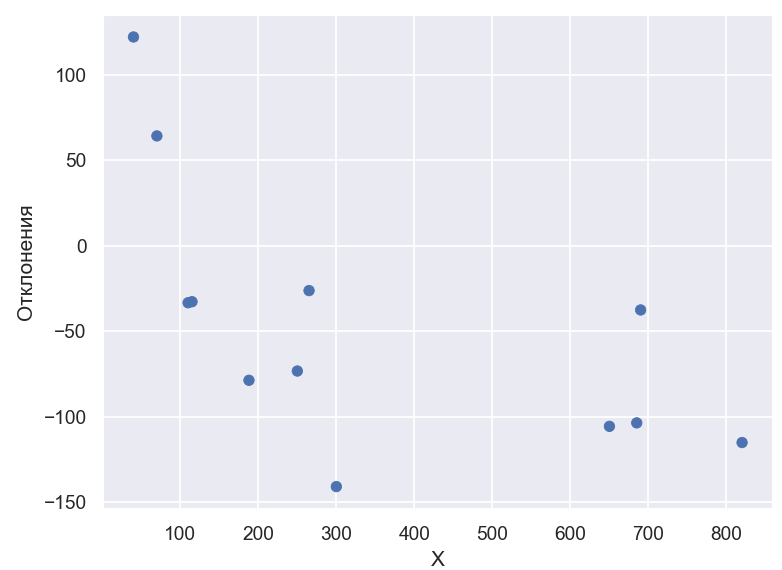


Рисунок 16 – График отклонения предсказанного значения от действительного.

На рисунке 16 показан график отклонения значений, где в качестве оси X взяты действительные значения веса, а на оси Y отклонение предсказанных значений от действительных в граммах. Как можно заметить, при малом весе отклонение выше, чем при более крупном весе.

По построенной модели линейной регрессии можно увидеть, что средняя ошибка аппроксимации (отличие действительных значений от предсказанных) составляет 52,88%, что на 1,05% лучше, чем у модели, построенной по двум параметрам. Однако отклонение более чем на 50% все ещё можно считать плохим результатом.

1. Удаление аномальных и пропущенных значений и выявление качественных изменений в модели.

Как было выявлено ранее, в наборе данных отсутствуют аномальные и пропущенные значения, а значит очистка данных не представляется возможной.

Часть 2

В данной части работа выполняется с набором данных database.csv, в которых хранятся данные об отчетах об убийствах в Соединенных Штатах за период 1980 - 2014.

1. Вывести датасет и типы его переменных

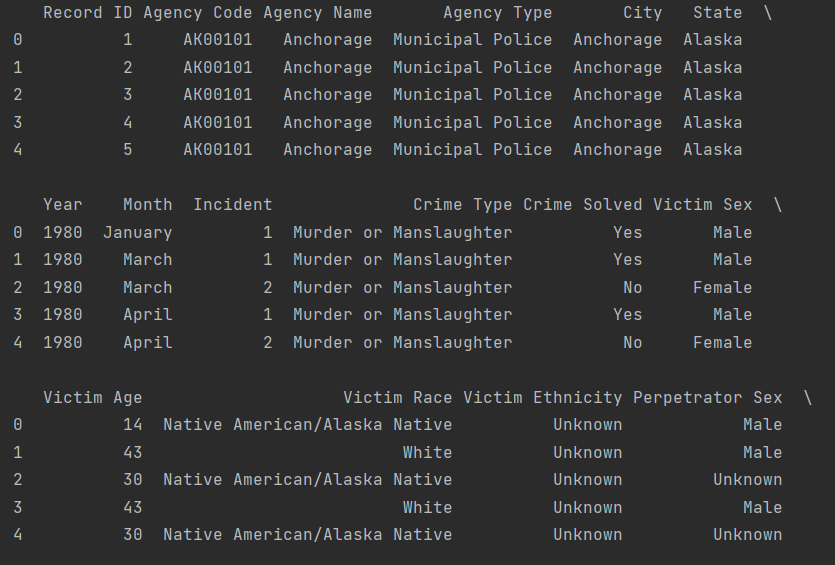


Рисунок 17.1 – Датасет database.csv

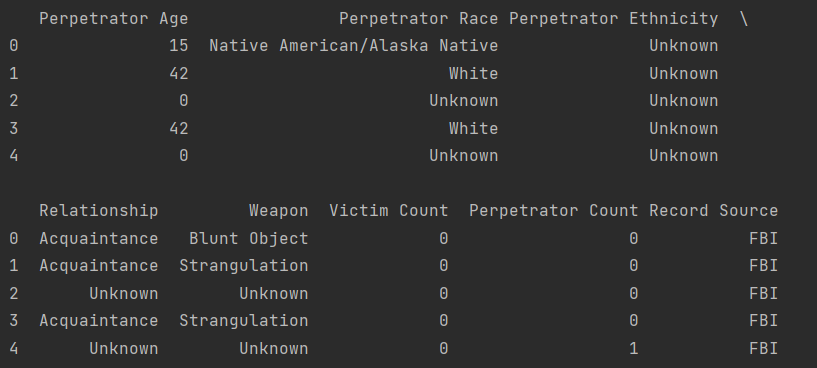


Рисунок 17.2 – Датасет database.csv

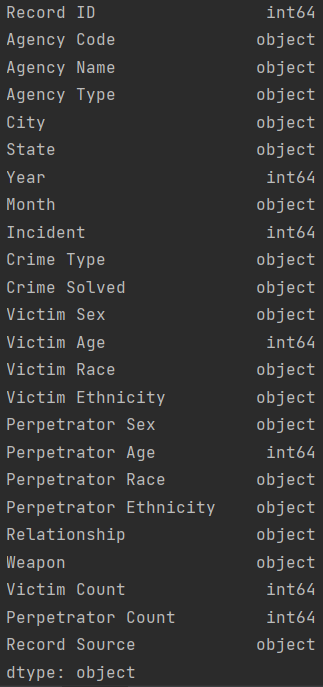
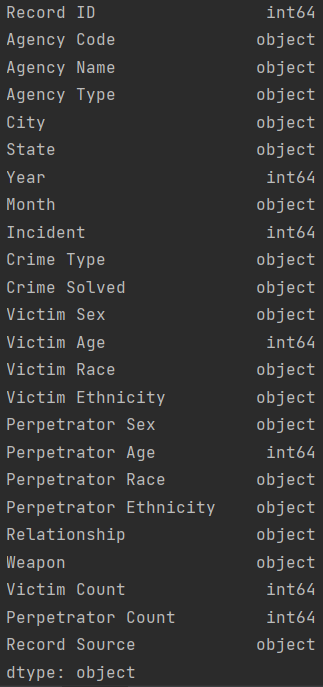


Рисунок 17.4 – Датасет database.csv

Данный датасет состоит из 24 переменных:

* Record ID – Идентификатор записи преступления
* Agency Code – Код агентства, которое занималось расследование преступления
* Agency Name – Название агентства, которое занималось расследование преступления
* Agency Type – Тип агентства, которое занималось расследование преступления
* City – Город, где было совершенно преступление
* State – Государство, где было совершенно преступление
* Year – Год, когда было совершенно преступление
* Month – Месяц, когда было совершенно преступление
* Incident – Какой именно инцидент произошёл
* Crime Type – Вид преступления
* Crime Solved – Завершенность раскрытия преступления
* Victim Sex – Пол жертвы
* Victim Age – Возраст жертвы
* Victim Race – Раса жертвы
* Victim Ethnicity – Этническая принадлежность жертвы
* Perpetrator Sex – Пол преступника
* Perpetrator Age – Возраст преступника
* Perpetrator Race – Раса преступника
* Perpetrator Ethnicity – Этническая принадлежность преступника
* Relationship – Отношение между жертвой и преступником
* Weapon – Оружие
* Victim Count – Количество жертв
* Perpetrator Count – Количество преступников
* Record Source – Источник записи

1. Используя визуализацию отобразить количественное распределение по различным показателям

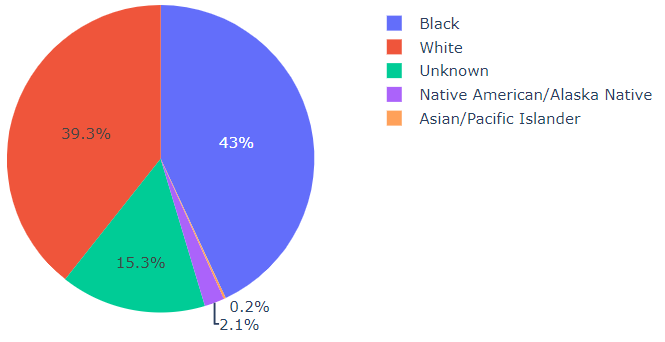
Первостепенно отобразим круговую диаграмму распределения по расе преступника. 

Рисунок 18.1 – Распределение расы преступников по количеству.

Как видно из круговой диаграммы, в нашем наборе данных присутствует 4 различных рас, а также неизвестные расы у нераскрытых преступлений. Больше всего (43%) чернокожих, далее (39,3%) идут белые люди, неизвестные расы (15,3%) занимают третью строчку, затем (2,1%) коренной американец/уроженец Аляски и наконец (0,2%) житель Азиатско/Тихоокеанских островов.

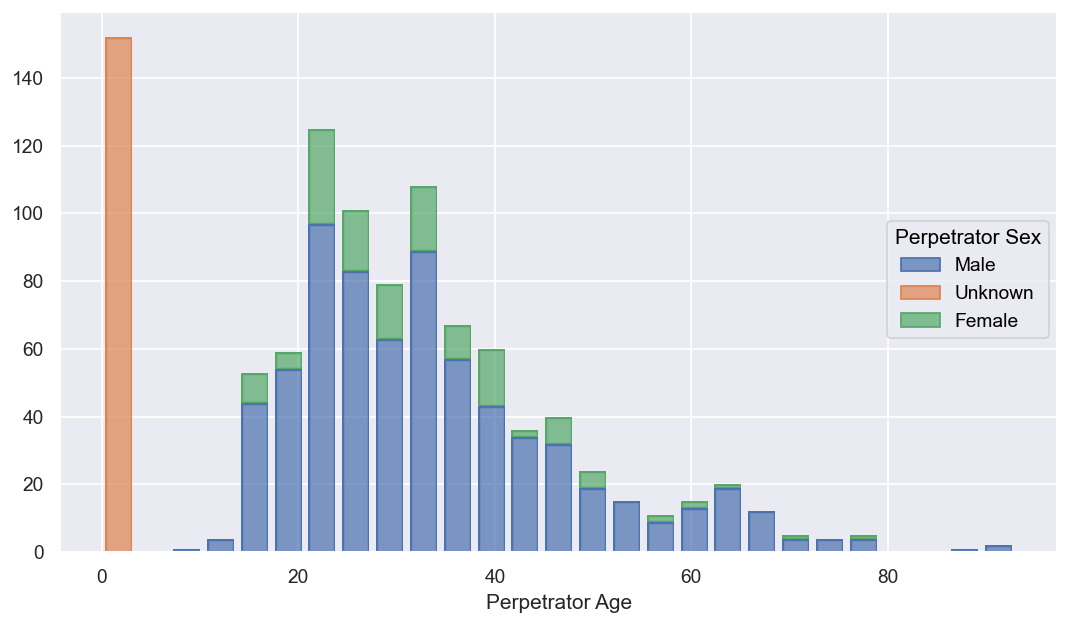


Рисунок 18.2 – Распределение количества преступников по возрасту.

Как видно из диаграммы наибольшее количество преступников имеет среднюю оценку между 20 и 35, в то время как преступников с возрастом ниже 15 и выше 80 практически нет. Также из данной диаграммы можно заметить, что в среднем мужчины совершают больше преступлений, что женщины.

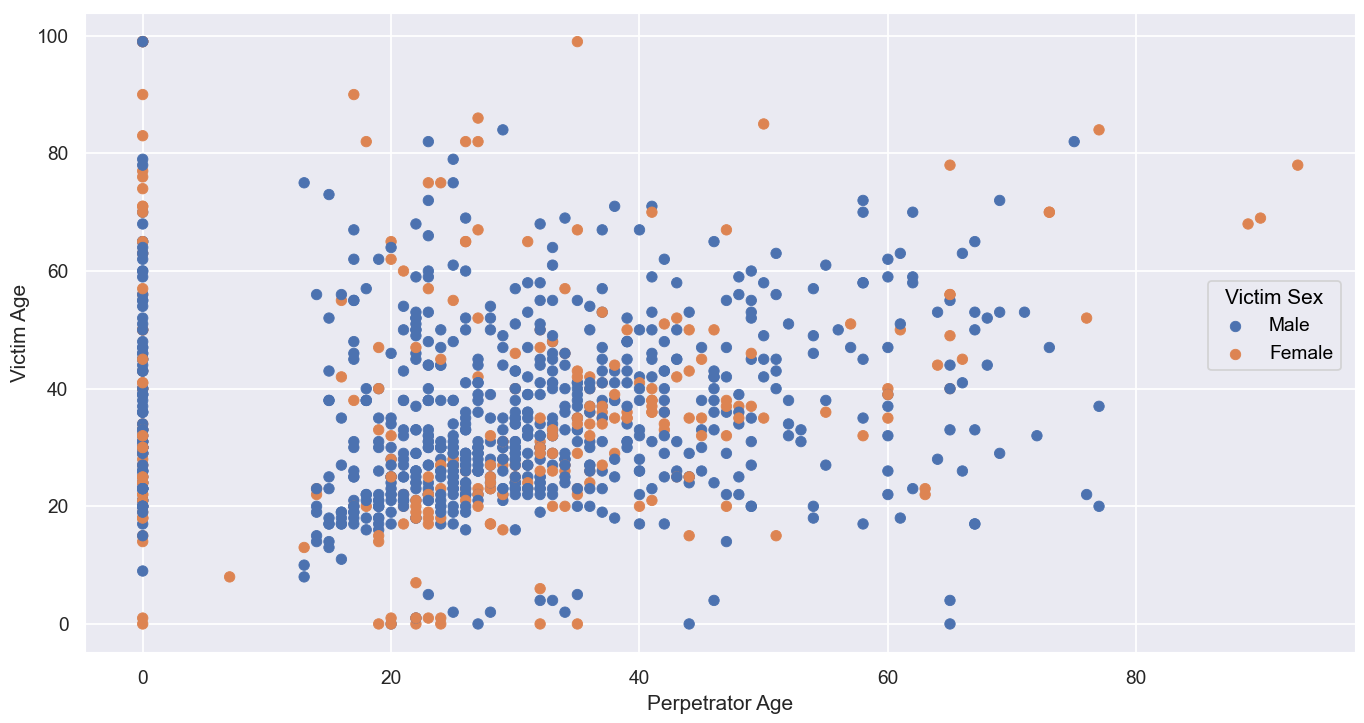


Рисунок 18.3 – Распределение возраста преступников по возрасту их жертв.

Как можно заметить из данного графика, что преступники в основном выбирают своих жертв из тех же возрастных категорий, что и они сами, причем пол жертв, так же, как и преступников, в большинстве случаев, является мужским.

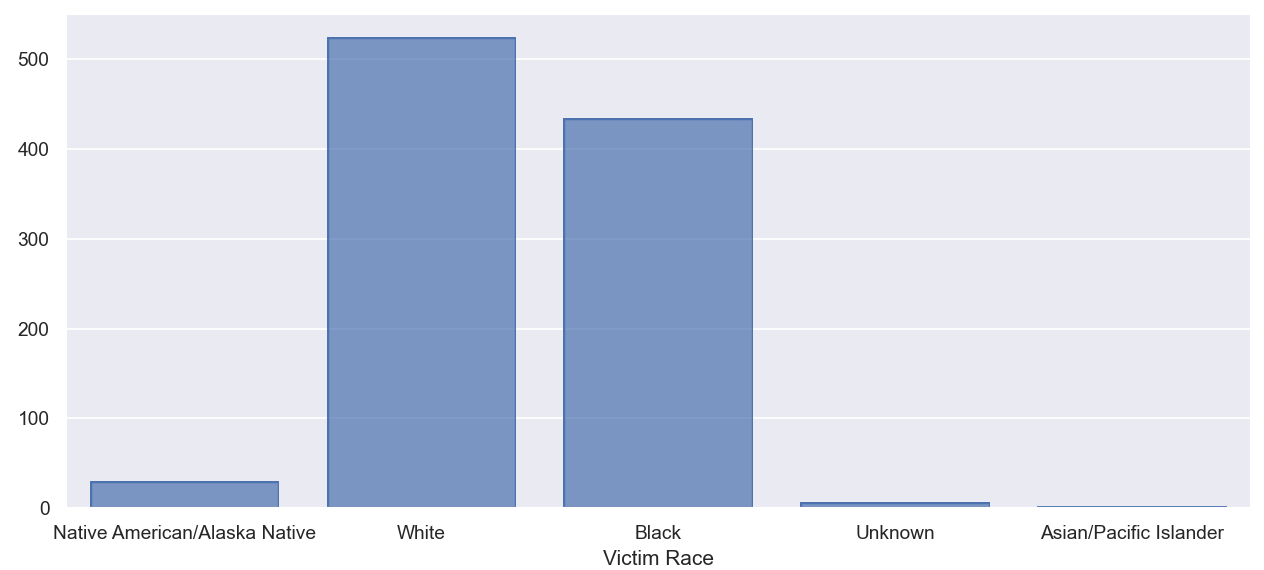


Рисунок 18.4 – Распределение расы жертв по их количеству.

Из данного графика видно, что наиболее частой расой жертвы является белые люди, потом идут темнокожие, а жертв, среди других рас, меньше 50.

1. Исследуйте есть ли какие-нибудь зависимые параметры.

Рассмотрим рисунок 18.3, где показано распределение возраста преступников по возрасту их жертв, по которому можно сделать вывод что возраст жертвы и преступника чаще всего находиться в одной возрастной категории. Также можно исследовать зависимость возраста преступника от его расы.

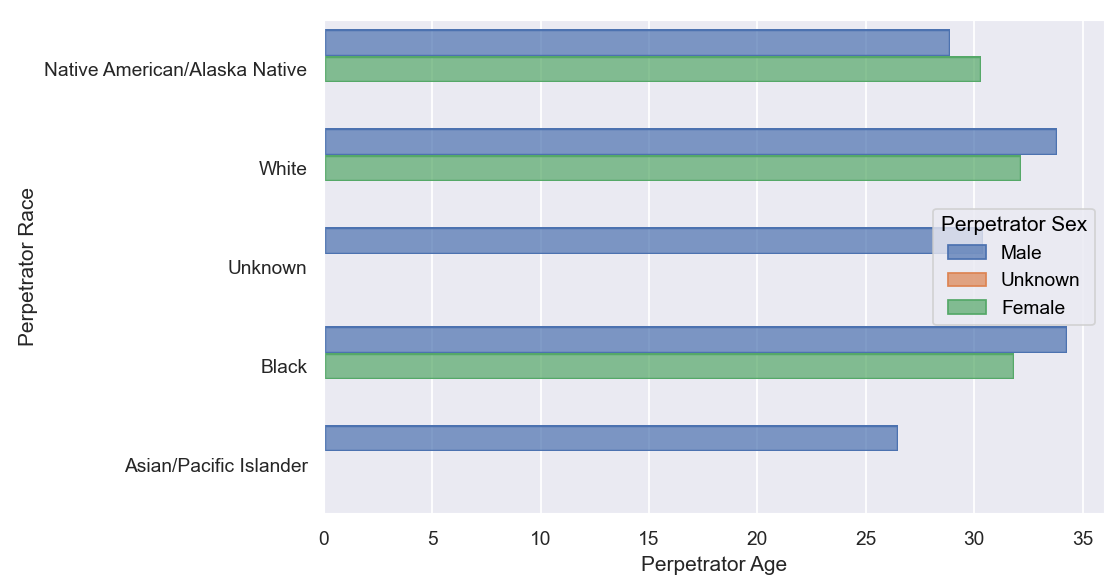


Рисунок 19 – График зависимости возраста преступника от его расы.

Как видно из графика, зависимость между возрастом и расой преступника отсутствует, так как в каждой расе присутствуют преступники одной возрастной группы.

1. Определите тип распределения количественных переменных

Рассмотрим рисунки 18.2 и 18.4 для определения типов распределения. Возраст преступника имеет нормальное распределение, в то время как распределение по расе преступника имеет распределение, где большая часть значений сосредоточена в середине рассматриваемого промежутка.

1. Существуют ли в этом наборе данных аномальные наблюдения?

В данном наборе данных отсутствуют аномальные наблюдения, так как люди совершают преступления в любом возрасте, даже такие тяжелые как убийства.

1. Проверьте данные на наличие пропущенных значений.

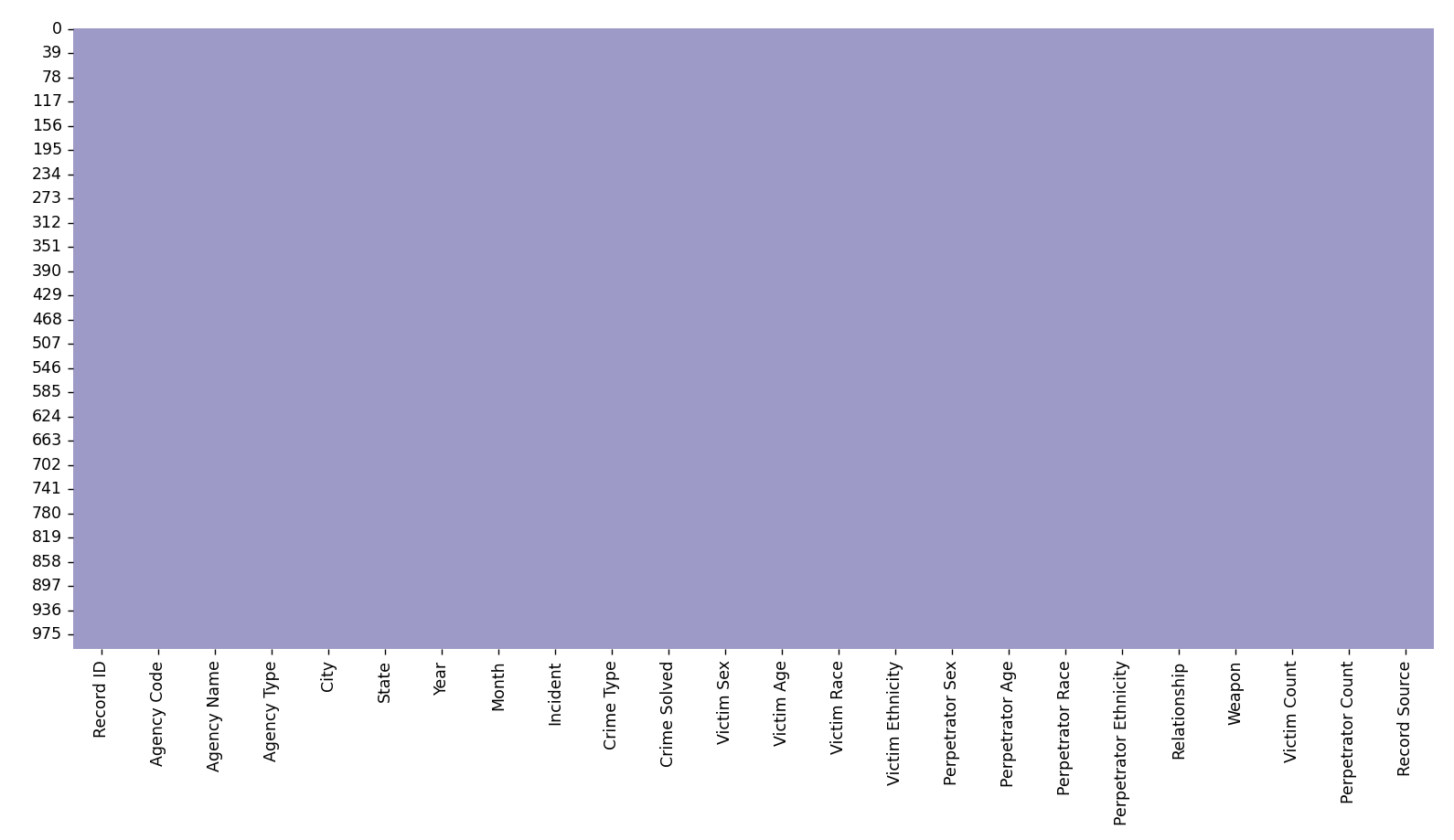


Рисунок 20 – График наличия пропущенных значений в датасете database.csv.

Из данного графика видно, что в данном датасете отсутствуют пропущенные значения

1. Реализовать визуализацию многомерных данных.

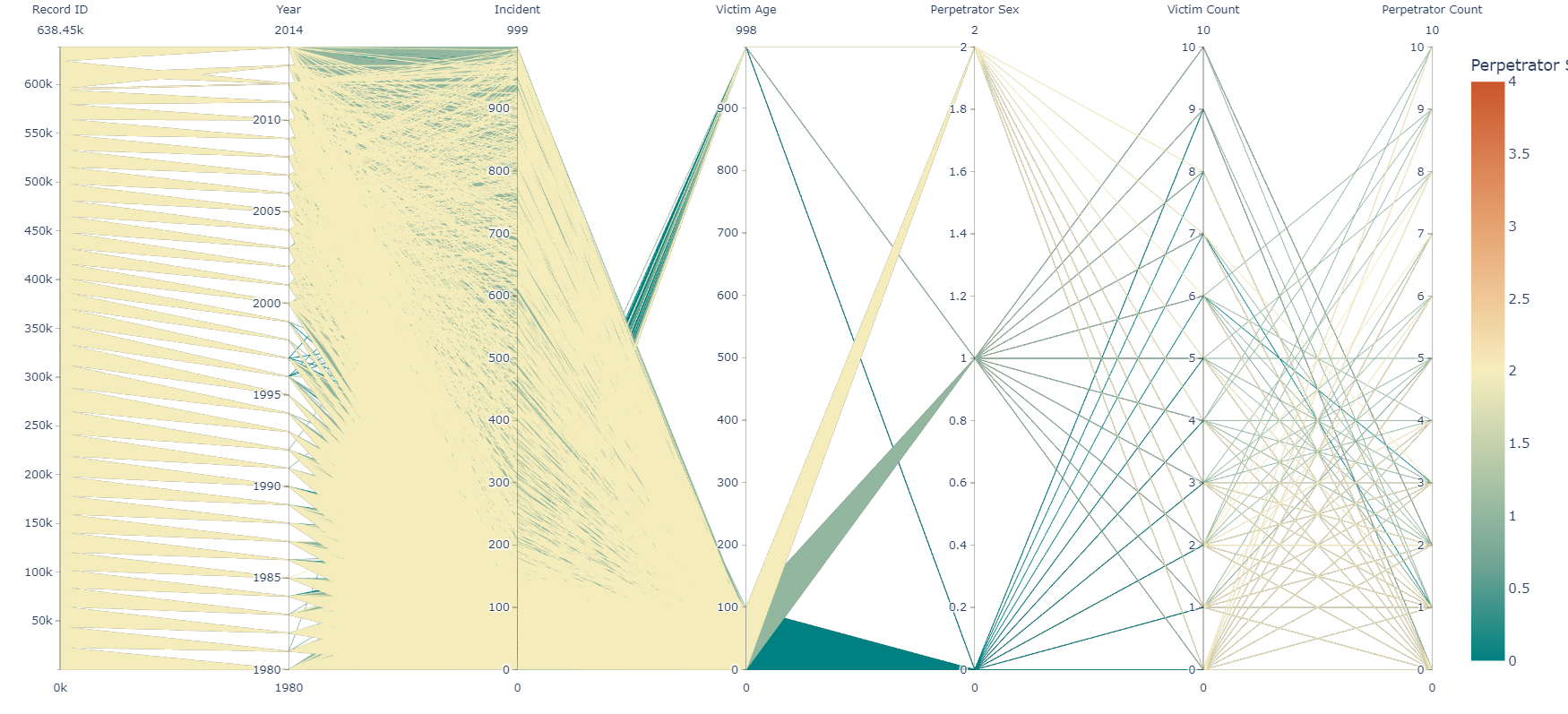


Рисунок 21 – Визуализация данных в виде параллельных координат.

Остальные типы графиков не представляется возможным построить из - за большого количества строк в наборе данных, а также малого количества численных переменных.

1. Реализовать модель регрессии для 2 параметров

В качестве параметров для построения модели будут взяты атрибуты датасета с данными об аниме: Victim Age(y) и Perpetrator Age(x). Также нецелесообразно рассматривать все расы преступников, так как ранее были выявлены различия в разных рас, поэтому для создания модели будут взять только чернокожих преступников. Построим диаграмму рассеивания для всех исследуемых пар значений.

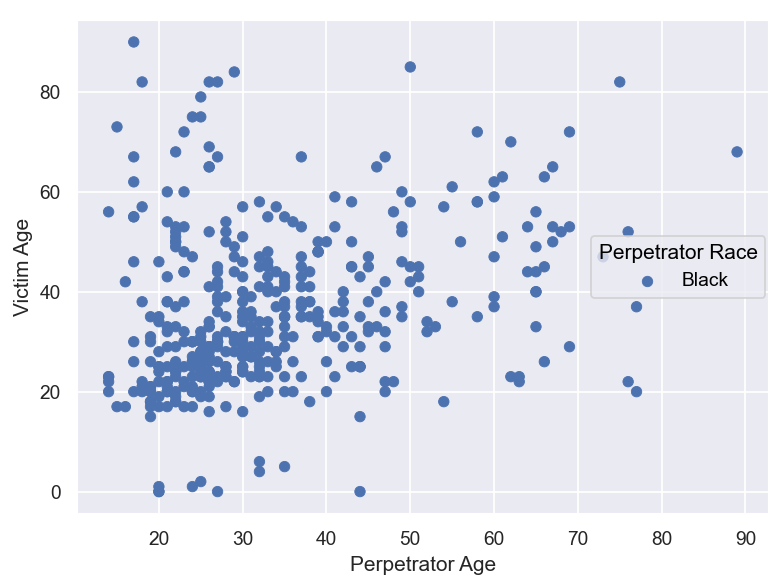


Рисунок 22 – Диаграмма рассеивания.

Как видно из диаграммы рассеивания, наиболее подходящей для нашего случая, является логарифмическая регрессия. Далее разделим все пары значений на тренировочную и тестовую выборку в соотношение 9:1, и определим коэффициенты уравнения регрессии.



Рисунок 23 – Коэффициент уравнения регрессии.

По полученным коэффициентам можно построить уравнение регрессии:   
**y = 0.3169 \* ln(x).**

Воспользуемся полученным уравнением для предсказания значений, и сравним полученные с действительными. Для этого воспользуемся тестовой выборкой.

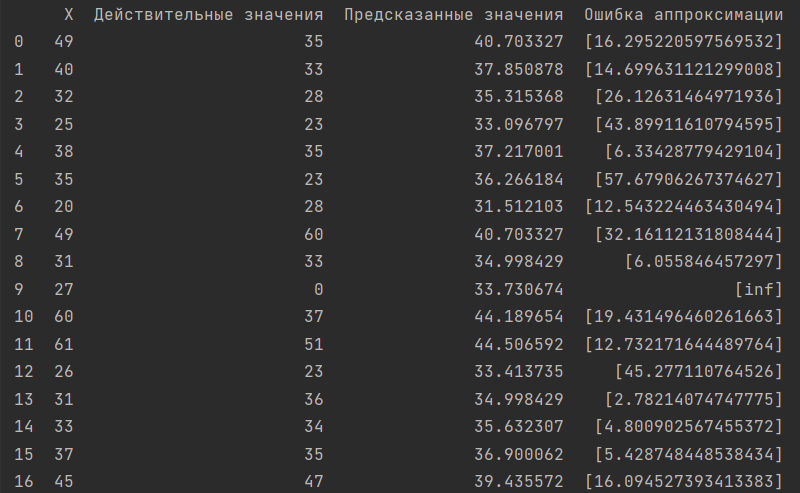


Рисунок 24 – Сравнение действительных и предсказанных значений.

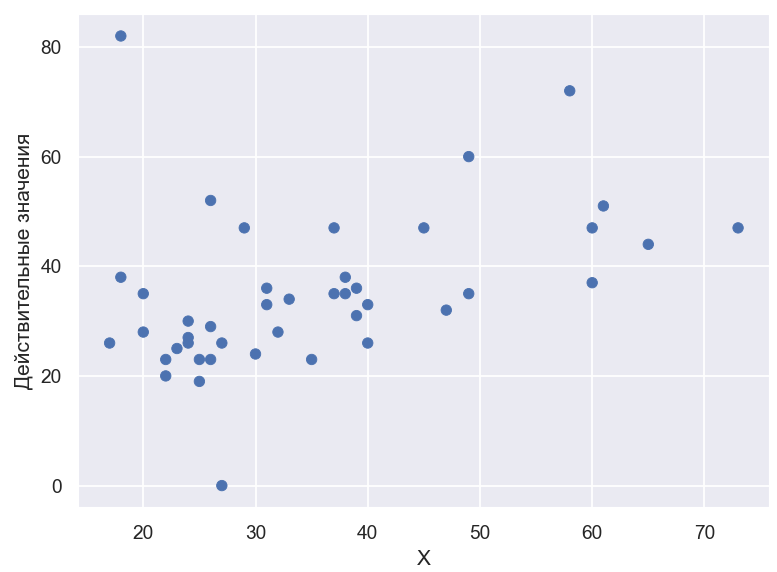
-

Рисунок 25 – График действительных значений тестовой выборки.

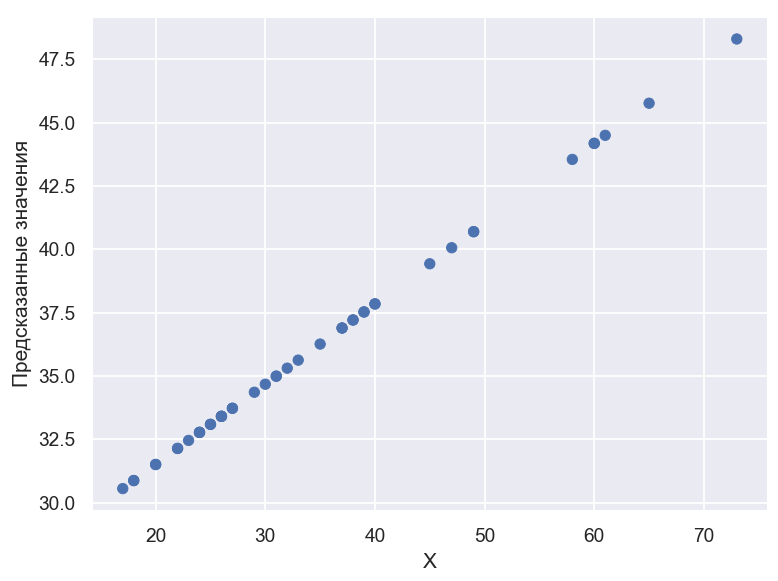


Рисунок 26 – График предсказанных значений тестовой выборки.

По полученным результатам можно увидеть, что средняя ошибка аппроксимации (отличие действительных значений от предсказанных) составляет 24,198%, что является неплохим результатом. Наибольшие отличия наблюдаются при расчёте возраста меньше 15 лет.

1. Построение модели регрессии по нескольким параметрам

Выполнение данной задачи не представляется возможным, так как в наборе данных отсутствуют какие - либо другие числовые параметры, которые могут влиять на возраст жертвы.