Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение   
высшего образования

**Иркутский национальный исследовательский технический университет**

|  |
| --- |
| Институт информационных технологий и анализа данных |
| Центр программной инженерии |

|  |
| --- |
|  |
| Отчет по |
| **Лабораторной работе №1**  **Подготовка данных для анализа** |
| наименование темы |

ЛАБОРАТОРНАЯ РАБОТА

по дисциплине

|  |
| --- |
| Методы анализа данных |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил студент группы |  | ИСТб-21-1 |  |  |  | Д.И. Морозов |
|  |  | Шифр группы |  | подпись |  | И. О. Фамилия |
| Проверил |  | доцент |  |  |  | Е. А. Осипова |
|  |  | должность |  | подпись |  | И. О. Фамилия |

Иркутск 2023 г.

Постановка задачи

В ходе лабораторной работы требуется, определившись с набором данных, выполнить подготовку данных к анализу с помощью языка программирования Python. Определить типы шкал, группирующие переменные и переменные отклика. Проверить набор данных на наличие пропущенных значений и, при наличии, исправить дата сет. Определить наличие выбросов с помощью диаграмм Кливленда, размаха и рассеяния.

Описание исходных данных

В качестве набора данных была выбрана статистика по ценам на Биткоин в период с 2010 по 2020 гг. В набор включена информация о дате, цене на открытии, высшей цене, низшей цене, цене на закрытии, cкорректированной цене закрытия и объёме торгов.

Анализ шкал и определение переменных

Таблица 1 – Типы шкал

|  |  |  |  |
| --- | --- | --- | --- |
| **Исходное имя** | **Перевод** | **Тип шкалы** | **Тип данных** |
| Overall rank\_2019 | Место в таблице | Рейтенговая | Числовой |
| Country or region | Цена на открытии | Объектная | Текстовый |
| Score\_2019 | Счёт за 2019 год | Количественная | Дробные числа |
| Rank\_2018 | Ранг за 2018 год | Количественная | Дробные числа |
| Score\_2018 | Счёт за 2018 год | Количественная | Дробные числа |
| Rank\_2017 | Ранг за 2017 год | Количественная | Дробные числа |
| Score\_2017 | Счёт за 2017 год | Количественная | Дробные числа |
| Rank\_2016 | Счёт за 2016 год | Количественная | Дробные числа |
| Score\_2016 | Счёт за 2016 год | Количественная | Дробные числа |
| 2020.Rank | Счёт за 2020 год | Количественная | Дробные числа |
| 2020\_Score | Счёт за 2020 год | Количественная | Дробные числа |
| Rank\_2022 | Счёт за 2022 год | Количественная | Дробные числа |
| Score\_2022 | Счёт за 2022 год | Количественная | Дробные числа |
| 2021. score | Счёт за 2021 год | Количественная | Дробные числа |
| 2021.Rank | Счёт за 2021 год | Количественная | Дробные числа |

Изучив набор данных, можно утвердить, что признак “Overall rank\_2019” является ключевым и позиция в этом признаке зависит от данных других признаков, а признак “ Country or region ” является вспомогательным и от него не зависит ни один другой признак, но без него не будет понятна сам датасет.

Определение пропущенных значений

Для выполнения лабораторной работы использовалась среда Google Colab. Для поиска пропущенных значений сначала было решено проверить соответствие предполагаемых типов строк и тому, какой тип данных они имеют в дата-фрейме. Далее представлены скриншоты с кодом и выводе среды разработки.

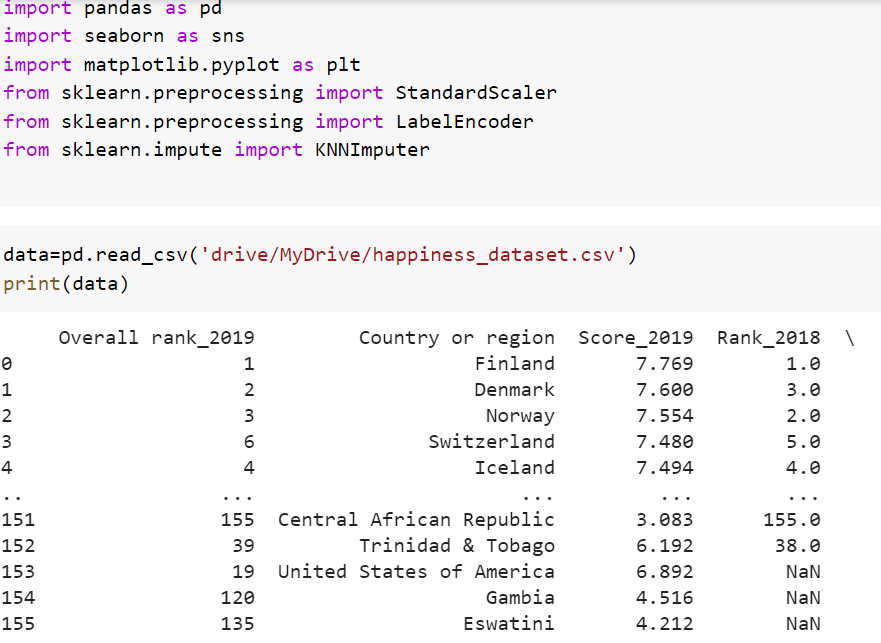


Рисунок 1 – Загрузка набора данных

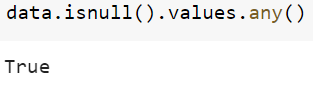


Рисунок 2 – Проверка на пустые значения

Пустые значения присутствуют, следовательно нам надо избавиться от них

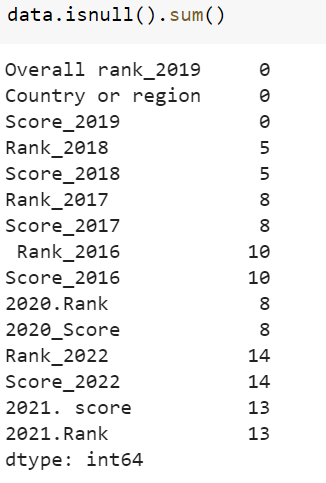


Рисунок 3 – Проверяем в каких столбцах сколько значений пропущено

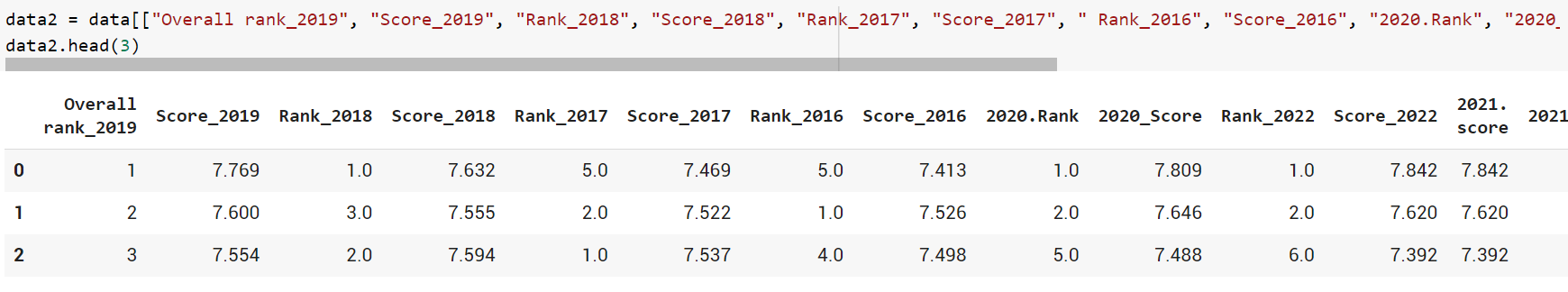


Рисунок 4 – Создание новой таблицы без строкового типа данных

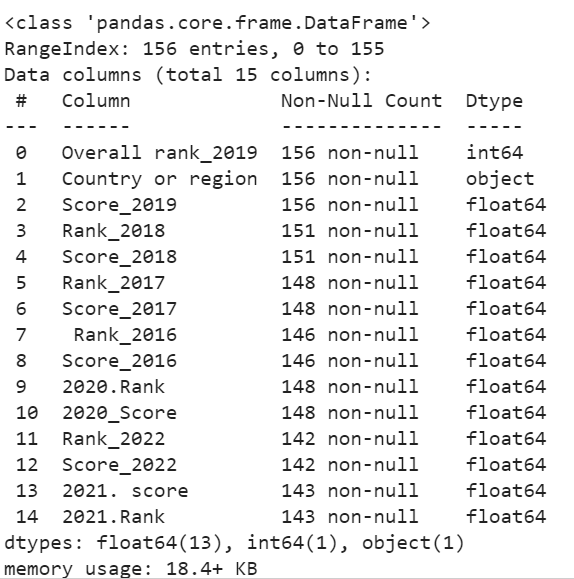


Рисунок 5 – Проверка строковых значений

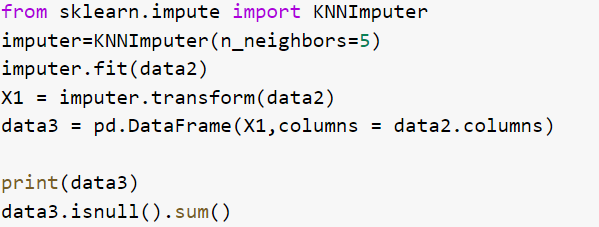


Рисунок 5 – Код импутации данных методом k-ближайших соседей

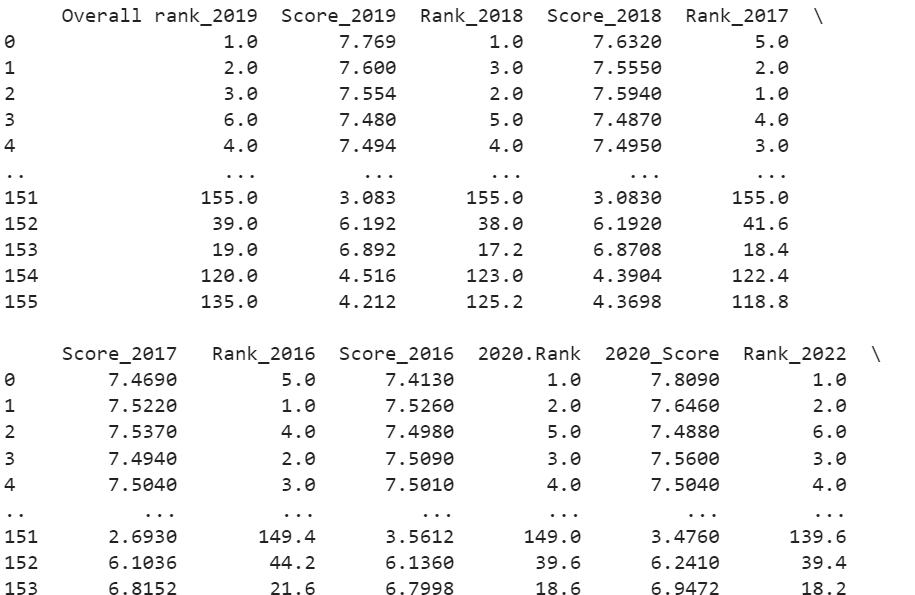


Рисунок 6 – Таблица без пустых значений

Итак, нам удалось выяснить, что во всех колонках отсутствуют пустые значения, а также все типы данных совпадают, можно переходить к построению диаграмм.

Построение диаграмм Кливленда

Для построения диаграммы Кливленда возьмем пару количественных признаков – счёт за 2018 год и за 2019 год.

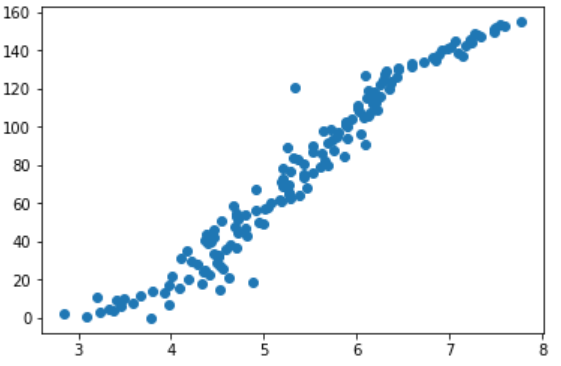


Рисунок 6 – Диаграмма Кливленда и возможные выбросы

Судя по диаграмме, набор не имеет достаточно большое количество выбросов, однако, исходя из области исследования, можно судить о рейтинге стран и их уровне счастья.

Диаграммы размаха

Для проверки значений, полученных в предыдущем пункте, проверим их, используя диаграмму размаха

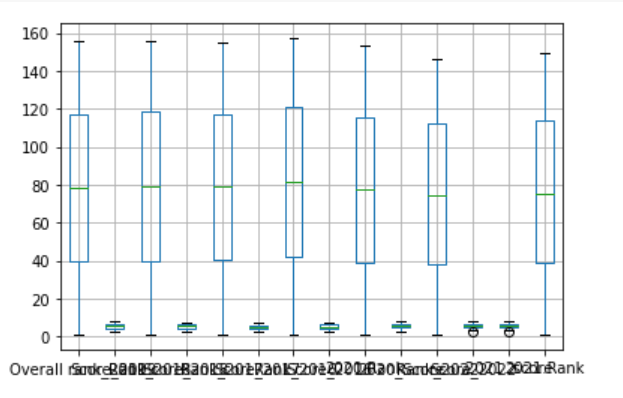


Рисунок 7 – Диаграмма размаха до нормализации данных

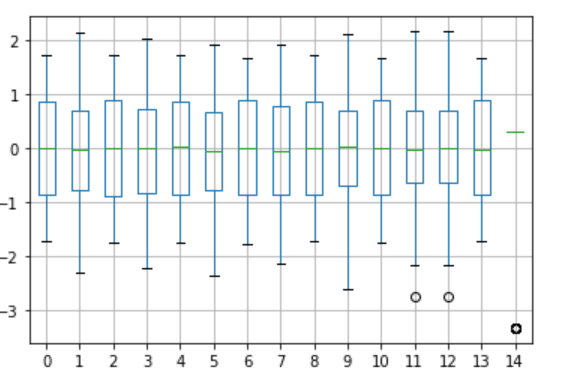


Рисунок 8 – Диаграмма размаха после нормализации данных

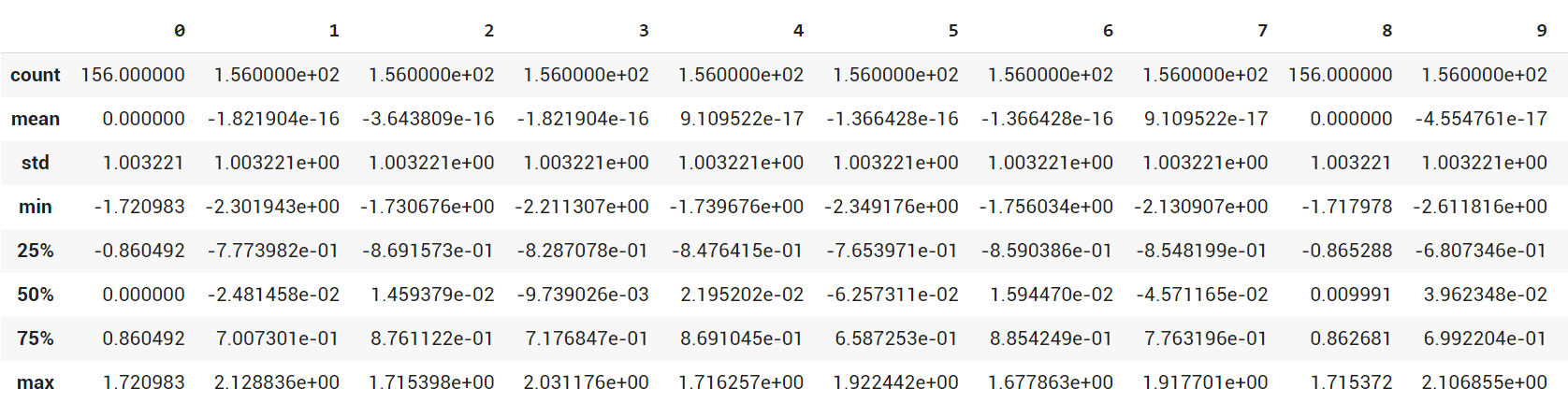


Рисунок 9 – Таблица после нормализации данных

В таблице (Рисунок 9) мы видим, что значения в некоторых строчках стали равны единице, а в некоторых почти равны нулю. Следовательно нормализация данных прошла успешно

Диаграммы рассеяния

Построим диаграмму рассеяния.

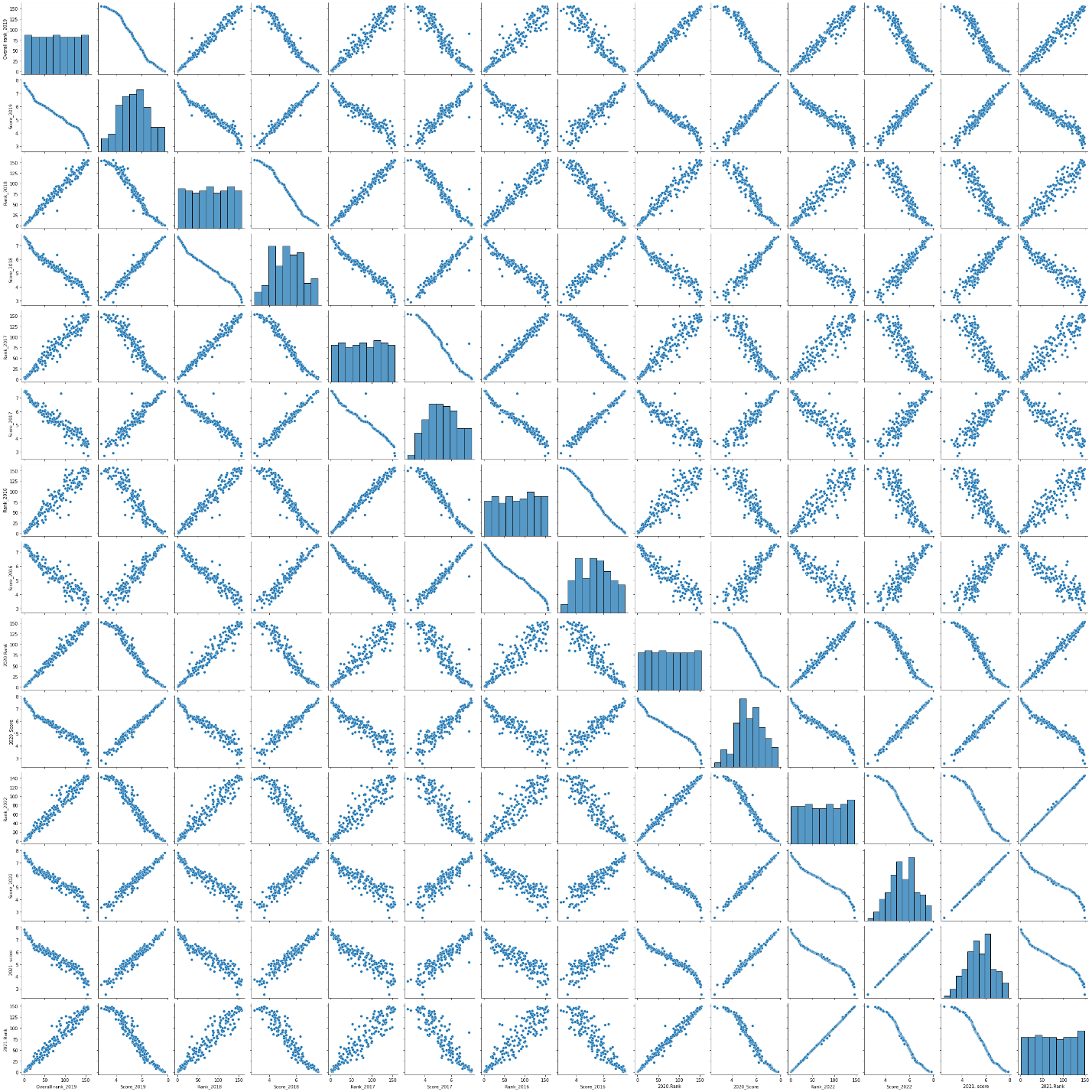


Рисунок 10 – Диаграмма рассеяния

По этой диаграмме можно сделать вывод, что рейтинг стран в разные годы почти не меняется.

Метод Forest

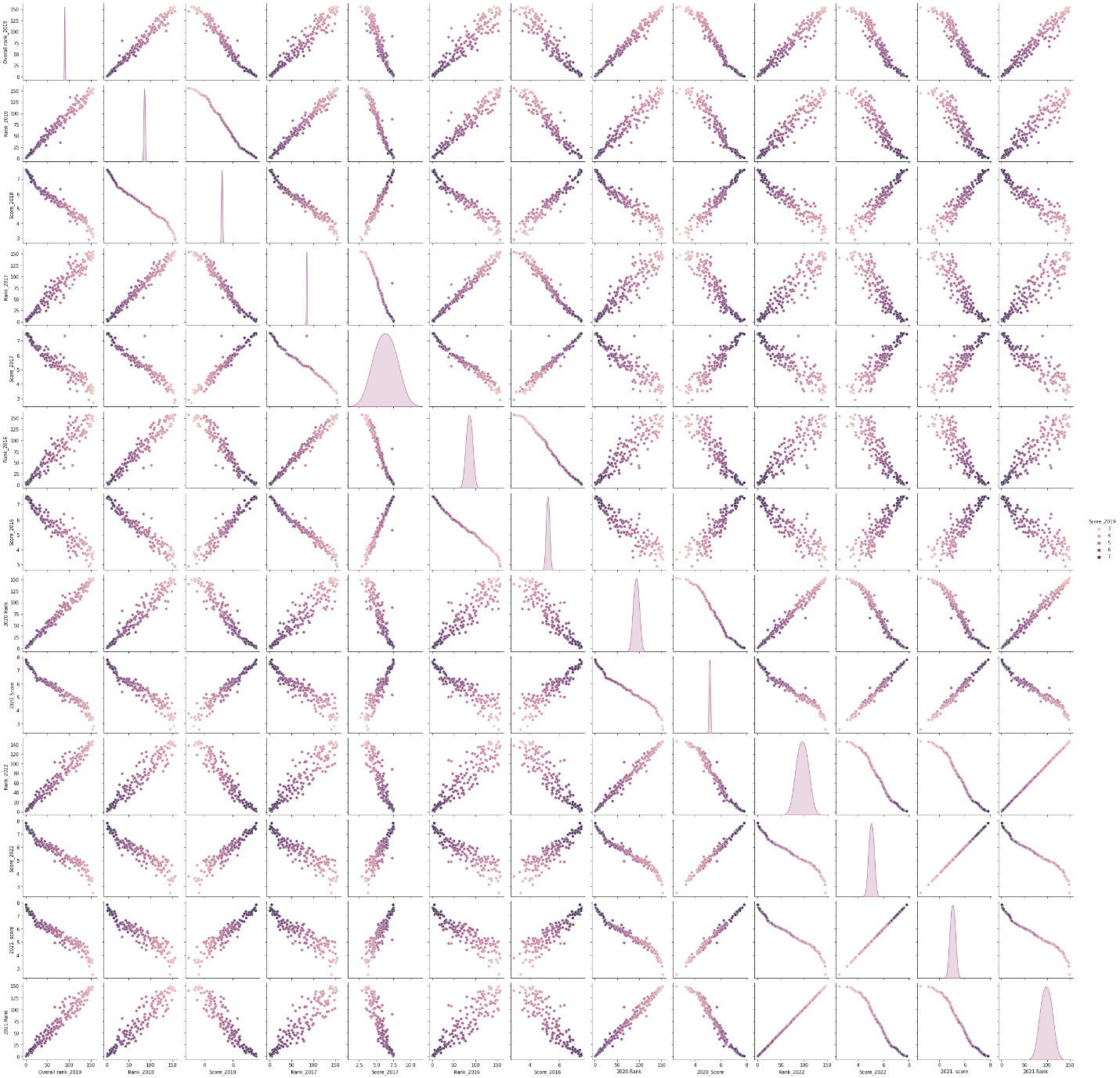


Рисунок 11 – Диаграмма рассеяния Forest

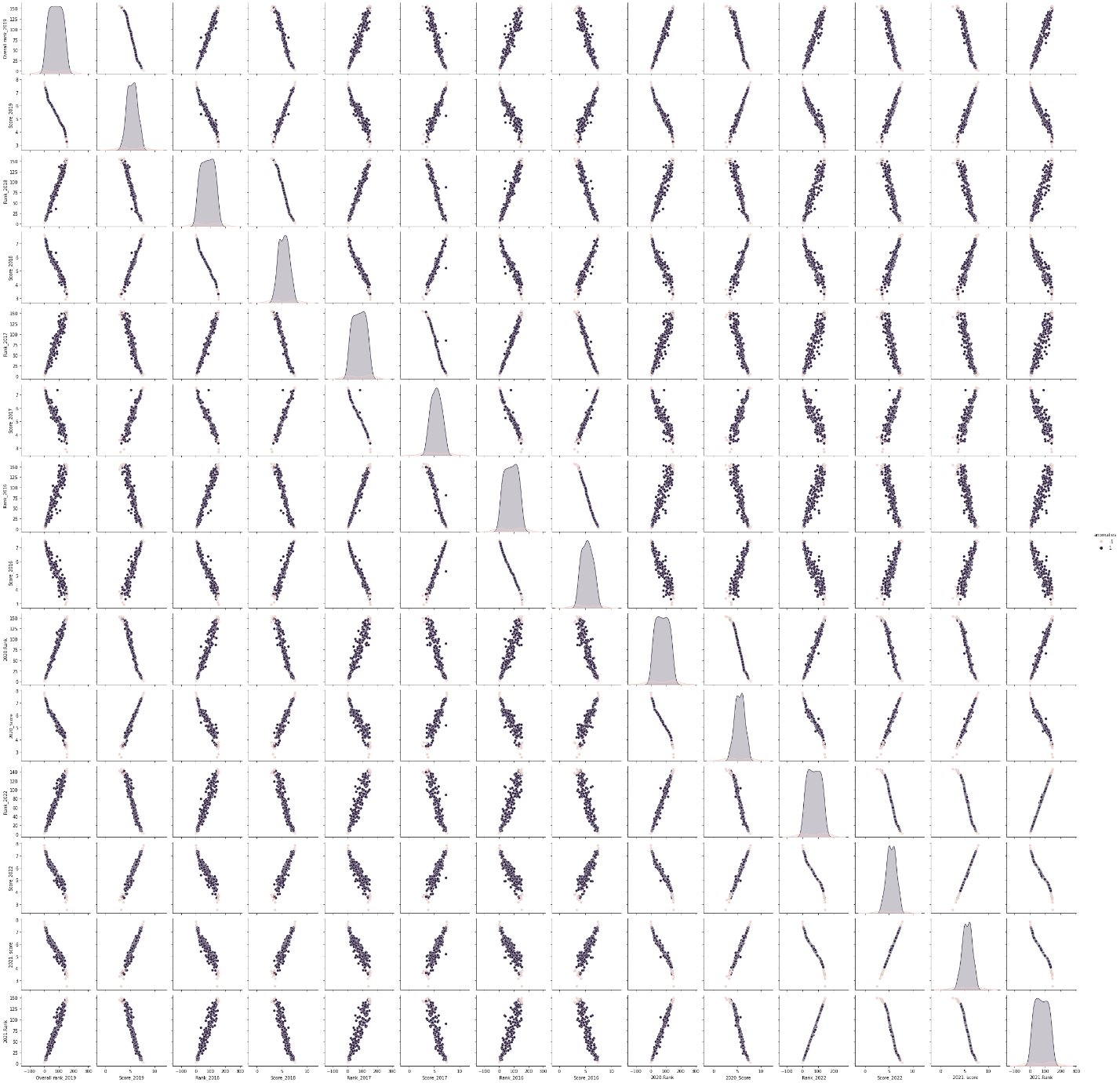


Рисунок 12 – Диаграмма рассеяния Forest после удаления выбросов

При просмотре диаграммы мы можем сделать вывод, что выбросы были удалены успешно

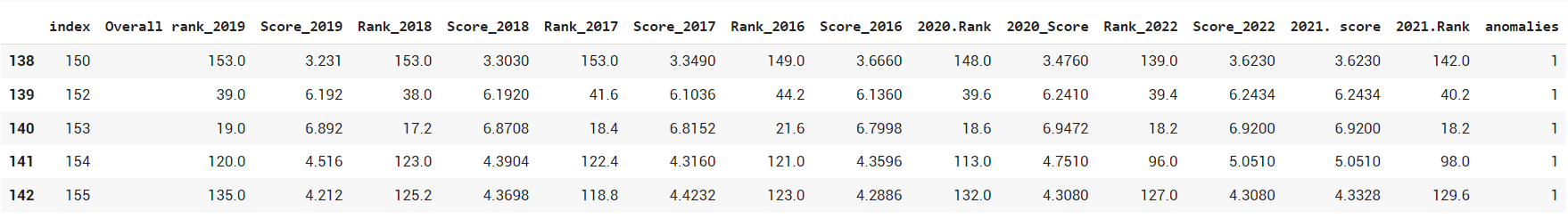


Рисунок 13 – Проверка таблицы после удалённых данных

Вывод

В результате выполнения лабораторной работы удалось применить на практике знания по нахождению пустых значений в наборе данных, анализе набора на выбросы. Удалось опробовать новые для себя инструменты, в частности, среду программирования Google Colab.