**Министерство науки высшего образования Российской Федерации**

**федеральное государственное автономное образовательное учреждение высшего образования**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»**

**(Университет ИТМО)**

Факультеттехнологий искусственного интеллекта

**Лабораторная работа №4**

Анализ данных на предмет выбросов и дрифтов

**Выполнили:**

Гончаренко Данила Олегович, группа J4150

Стрельницкая Татьяна Викторовна, группа J4140

**Преподаватель:**

Старобыховская Анастасия Александровна

Санкт-Петербург

2024

Оглавление

[Задание 2](#_Toc180409830)

[Основные этапы 3](#_Toc180409831)

[Вывод 8](#_Toc180409832)

# Задание

В рамках данной лабораторной работы поставлена цель – проанализировать датасет предыдущих лабораьорных работ на предмет шумов, отклонений и дрифтов.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Взять задачу из прошлых лабораторных и датасет,

2. Проанализировать данные для датасета на наличие шумов, выбросов/аномалий, дрифтов.

3. Выбрать алгоритм определения дрифта и применить его,

3+. Прикрутить инструмент мониторинга.

4. Проанализировать полученные результаты. Сделать выводы и описать почему они могли получиться именно такими.

6. Написать отчёт.

# Основные этапы

Была взята задача классификации качества вина.

Входные переменные - различные физико-химические свойства вина – представлены в таблице 1.

Таблица 1 – Набор входных данных для модели логистической регрессии

|  |  |
| --- | --- |
| **Название характеристики** | **Описание** |
| Fixed Acidity | Уровень кислотности, который остается после ферментации |
| Volatile Acidity | Количество уксусной кислоты, влияющей на вкус |
| Citric Acid | Лимонная кислота – усиливает вкус, придает свежесть |
| Residual Sugar | Сахар, остающийся после ферментации |
| Chlorides | Хлориды – содержание соли |
| Free Sulfur Dioxide | Свободный диоксид серы – действует как противомикробное средство |
| Total Sulfur Dioxide | Общее количество диоксида серы |
| Density | Плотность вина, связанная с содержанием алкоголя и сахара |
| pH | Уровень кислотности |
| Sulphates | Сульфаты, способствующие микробной стабильности |
| Alcohol | Процент алкоголя |
| Quality | Сенсорная оценка качества (от 0 до 10) |

Выходная переменная: показатель качества (0-10), определяемый на основе сенсорных данных.

Задача – определить выбросы и отклонения в датасете.

Взят датасет качества вина из прошлых лабораторных работ.

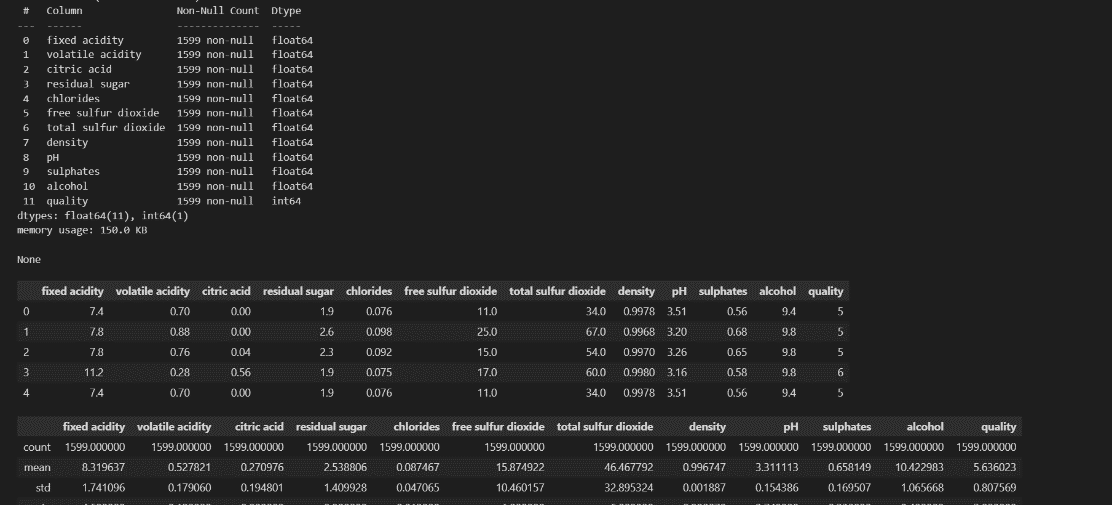


Рисунок 1 – Исходный датасет качества вина.

Далее отделяем зависимые переменные от целевой переменной. Для анализа были выделены следующие переменные: 'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol'.

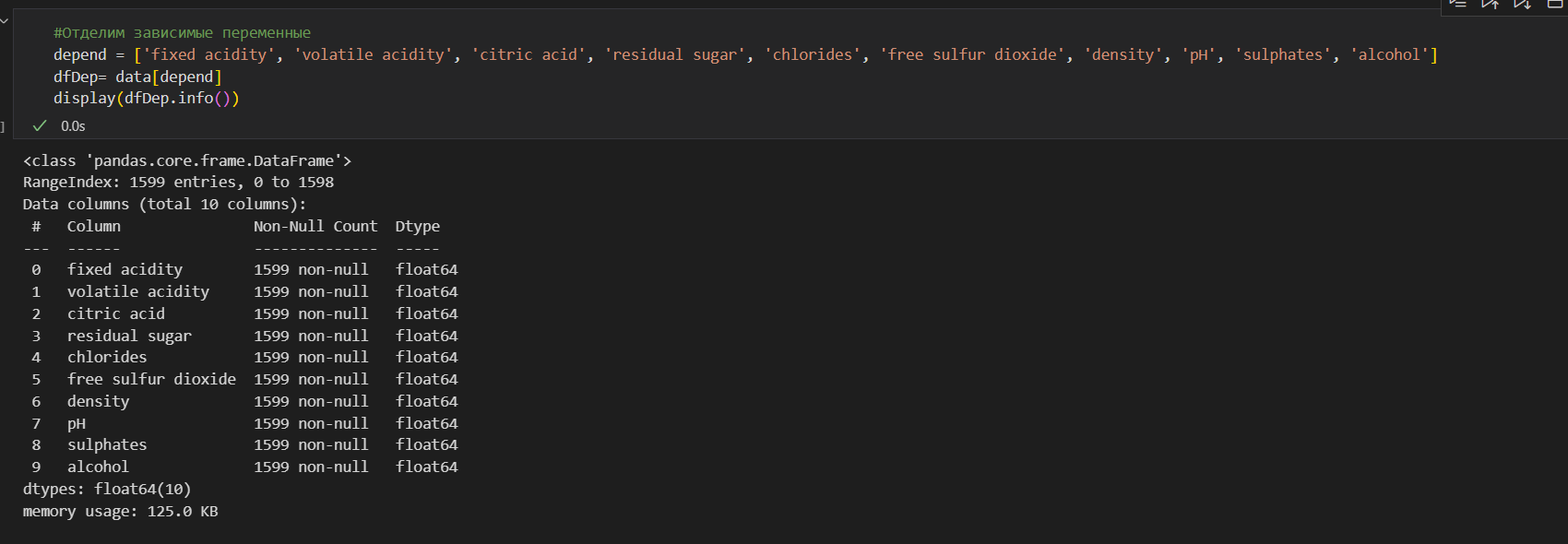


Рисунок 2 – Подготовленный датасет зависимых переменных

Проводим анализ выбросов при помощи ящиков с усами. При помощи ящиков с усами были выявлены выбросы по каждому показателю.

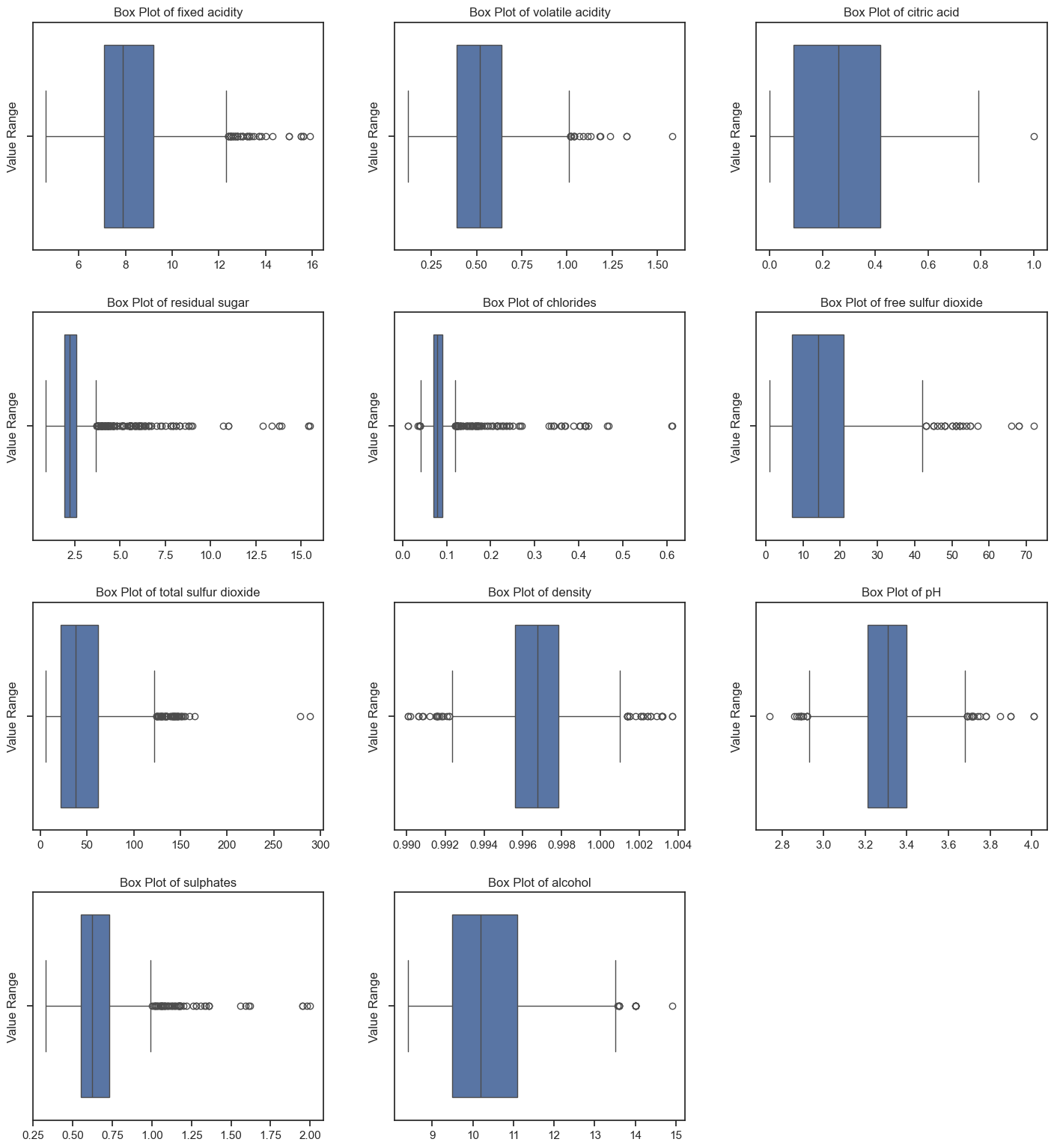


Рисунок 3 – Ящики с усами по зависимым переменным.

Затем для нахождения дрифта датасет был разделен на два тестовых датасата..

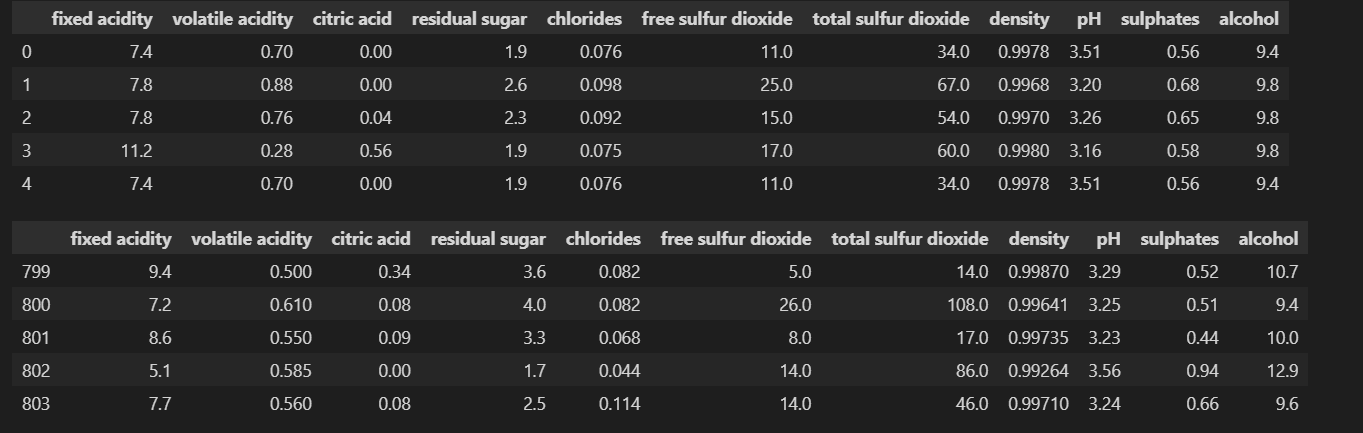


Рисунок 4 – Тестовые датасеты

Для нахождения дрифта данных был выбран алгоритм Колмагорова-Смирнова.

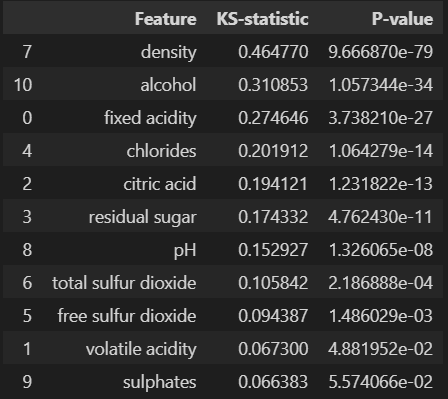


Рисунок 5 – Результаты работы алгоритма Колмагорова-Смирнова.

Для мониторинга был использован EVENDITLY AI. Для анализа дрифтов датасет был разделен на два равных тестовых сабсета. В результате анализа было найдено 7 колонок, имеющих дрифты: chlorides, alcohol, citric acid, pH, free sulfur dioxide, density.

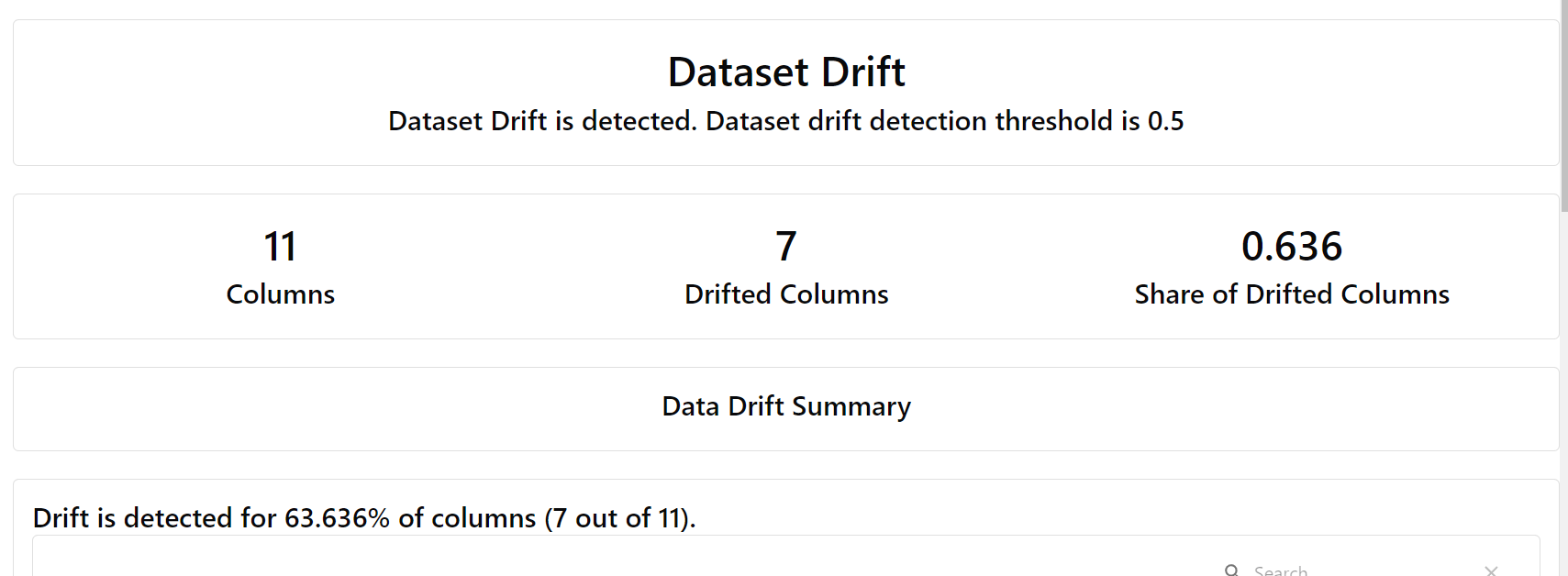


Рисунок 6 – Результаты анализа датасета на наличие дрифтов.



Рисунок 7 – Результаты анализа датасета по переменным.



Рисунок 8 – Data drift clorides.

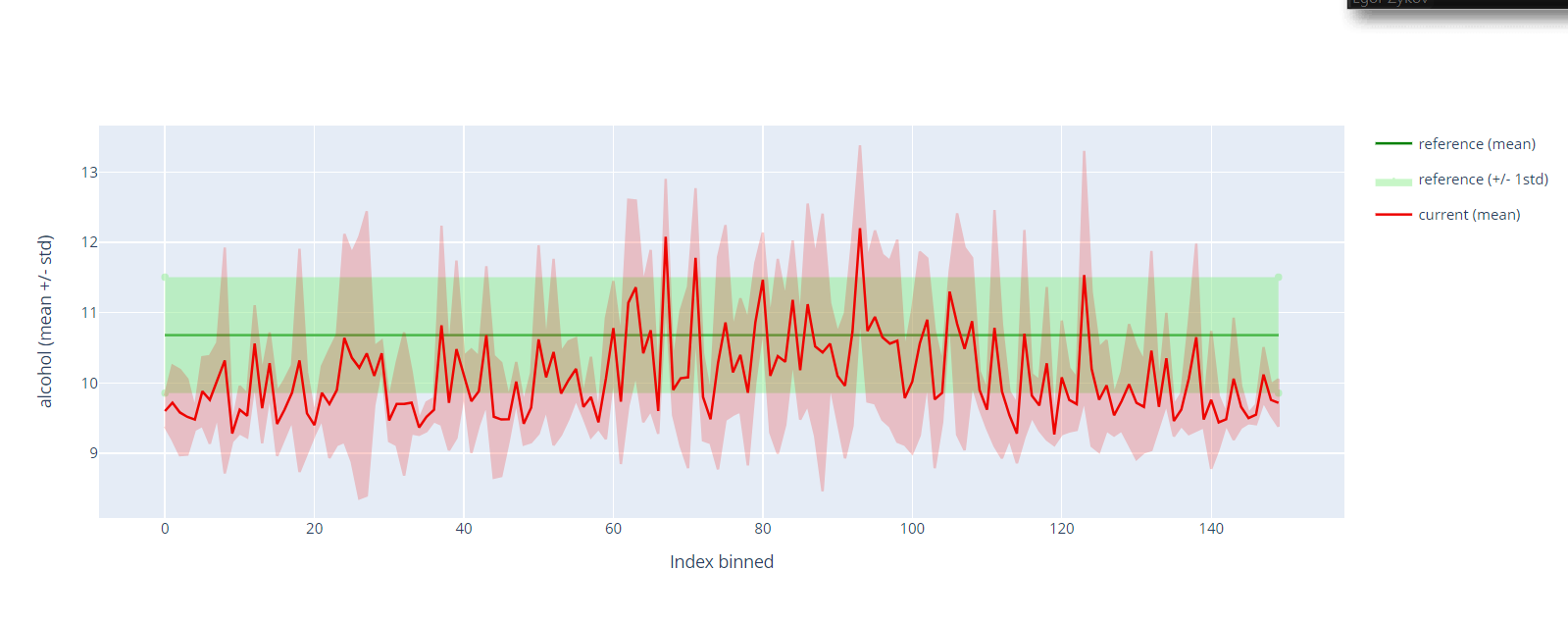


Рисунок 9 – Data drift alcohol

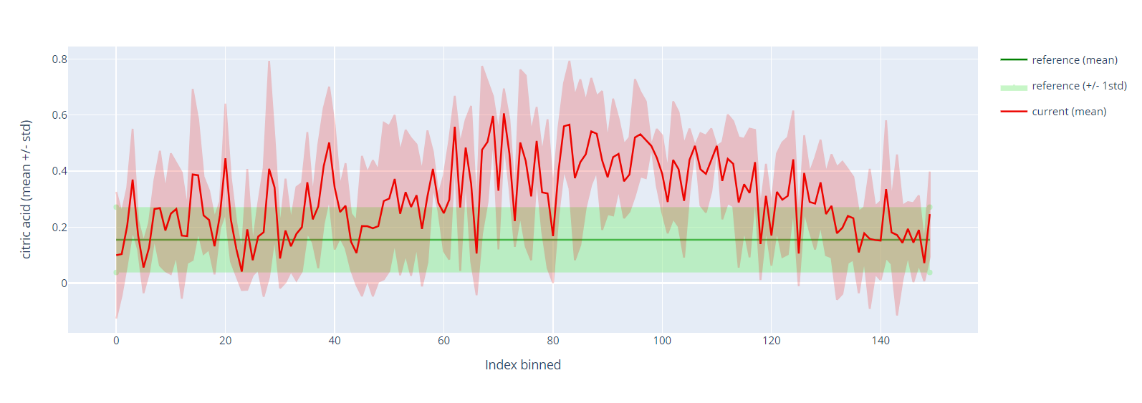


Рисунок 10 – Data drift citric acid



Рисунок 11 – Data drift pH



Рисунок 12 – Data drift free sulfur dioxide



Рисунок 13 – Data drift density

# Выводы

В результате определения дрифта при помощи алгоритма Колмагорова-Сморнова можно сделать вывод, что данные имеют не выбросы каждому показателю, а дрифты наблюдаются в 7 показателях. [Ссылка](https://colab.research.google.com/drive/1V7l23wugXPJqddGuqcxWMfedJ3DfxgWw?usp=sharing)