**МОЛДАВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ МАТЕМАТИКИ И ИНФОРМАТИКИ**

**ДЕПАРТАМЕНТ ИНФОРМАТИКИ**

Анализ и визуализация данных в сфере вин

**Индивидуальная работа номер 4**

Доцент университета: Поятэ Анатолий

Автор: студент группы IASD2401 (РУ), Жураковски Владислав

**Кишинёв, 2024**

СОДЕРЖАНИЕ

[I. Введение 3](#_Toc8156)

[II. Предобработка данных 3](#_Toc31986)

[2.1. Очистка данных 3](#_Toc9485)

[2.2. Преобразование данных 4](#_Toc20244)

[III. Исследовательский анализ данных 4](#_Toc2216)

[3.1. Статистические показатели 4](#_Toc6373)

[3.2. Анализ категориальных переменных 4](#_Toc17783)

[3.3. Предварительный анализ текста 5](#_Toc11089)

[3.4. Генерация облака слов 6](#_Toc27838)

[3.5. Корреляция слов 6](#_Toc28279)

[IV. Анализ корреляций 8](#_Toc31358)

[V. Разработка приложения Streamlit 10](#_Toc12947)

[VI. Результаты и выводы 12](#_Toc1407)

# I. Введение

**Цель** данной лабораторной работы заключается в применении методов анализа, предобработки и визуализации данных с использованием набора данных о винах. В рамках работы анализируются числовые, категориальные и текстовые данные для выявления зависимостей между характеристиками вин, а также создаются наглядные графики для интерпретации результатов.

**Набор данных** содержит следующую информацию о винах:

* **country**: Страна происхождения
* **description**: Текстовое описание вина
* **designation**: Специфическое наименование вина
* **points**: Баллы (рейтинг) вина
* **price**: Цена бутылки вина (в долларах США)
* **province**: Провинция происхождения вина
* **region\_1 и region\_2**: Регионы происхождения
* **variety**: Сорт винограда
* **winery**: Винодельня-производитель
* **title**: Полное наименование вина
* **vintage**: Год сбора урожая
* **alcohol**: Содержание алкоголя (%)
* **category**: Категория вина (например, красное, белое, игристое)

# II. Предобработка данных

## 2.1. Очистка данных

Были идентифицированы пропущенные значения в колонках **points**, **price** и **description**.

Для обработки пропущенных значений использовались следующие методы:

* + Для числовых данных (**points**, **price**) применялась **медиана**.
  + Для текстовых данных пропущенные значения заменялись **модой** (наиболее часто встречающимся значением).
* Были удалены дубликаты с помощью функции drop\_duplicates.

## 2.2. Преобразование данных

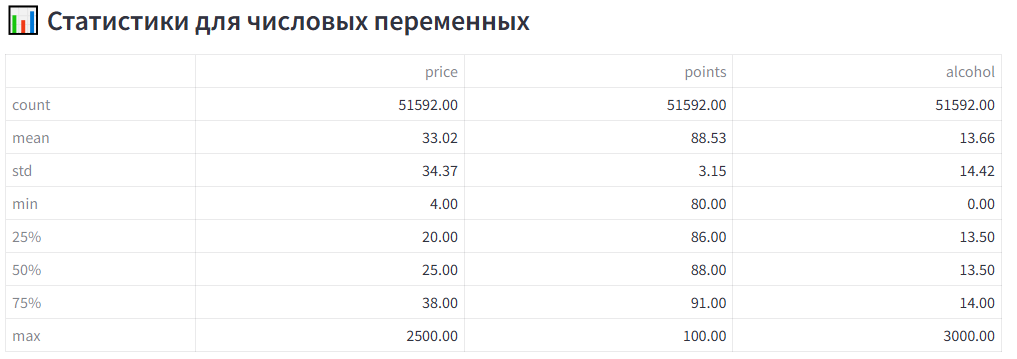
Колонка **points** была преобразована в числовой формат с помощью pd.to\_numeric.

Добавлена новая колонка price\_quality\_ratio, отражающая соотношение цена/качество: \text{price\_quality\_ratio} = \frac{\text{price}}{\text{points}}

# III. Исследовательский анализ данных

## 3.1. Статистические показатели

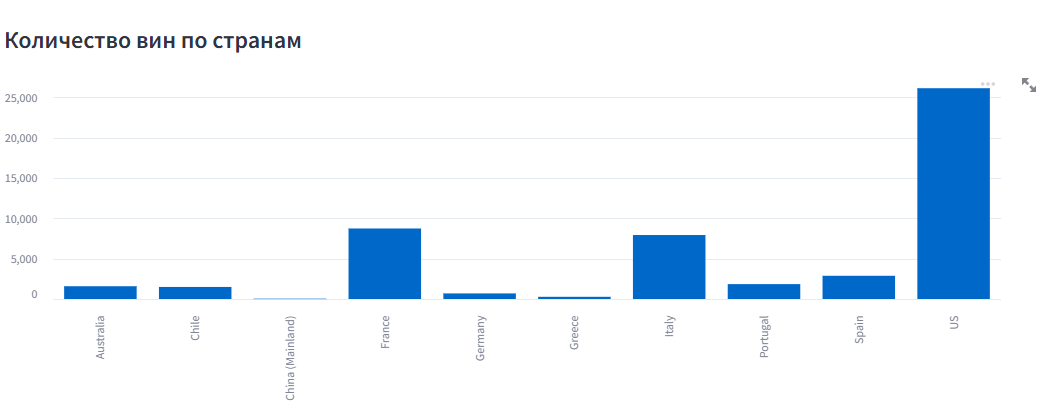
Для числовых колонок (**price**, **points**, **alcohol**) были рассчитаны статистические показатели: **среднее значение**, **медиана**, **стандартное отклонение**, **минимальные и максимальные значения**.



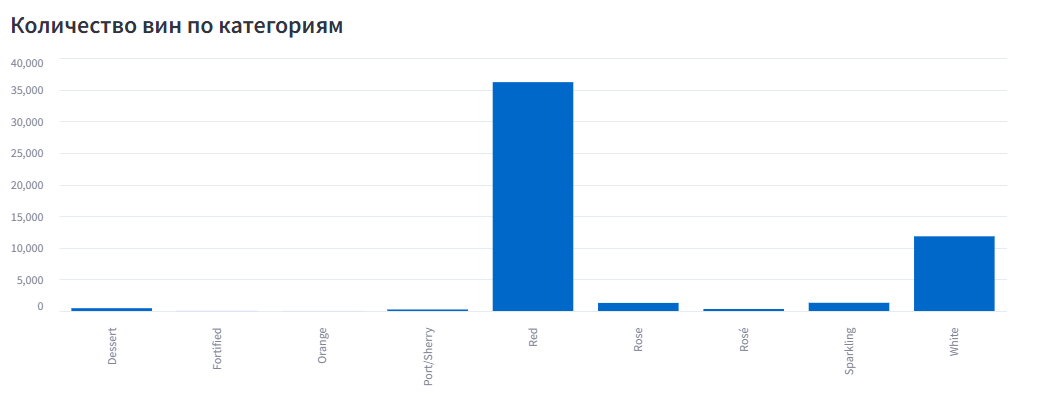
## 3.2. Анализ категориальных переменных

Были проанализированы:

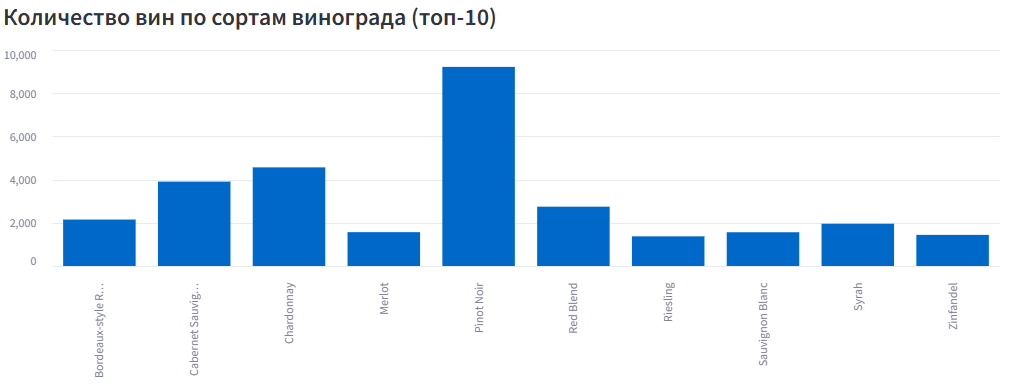
* **Распределение вин по странам**



* **Распределение вин по категориям (красное, белое и т.д.)**



* **Топ-10 сортов винограда по популярности**



## 3.3. Предварительный анализ текста

* Выделены **наиболее частотные слова** из колонки **description**, исключив общие слова ("и", "это", "с", и т.д.).
* Рассчитана **средняя длина описания** в словах.



## 3.4. Генерация облака слов

Было создано **облако слов** с помощью библиотеки WordCloud для визуализации наиболее часто встречающихся слов в описаниях вин.



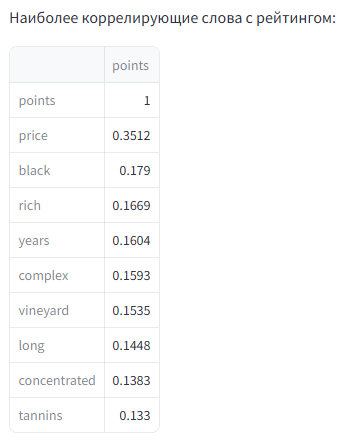
## 3.5. Корреляция слов

Проанализирована корреляция между частотными словами:

* **ценой вина (price)**

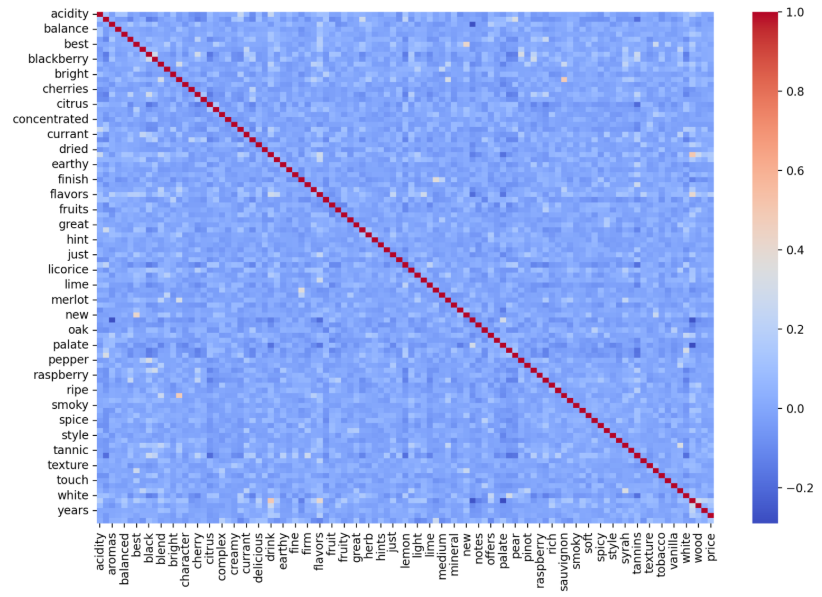


* **рейтингом (points)**



Корреляции были визуализированы с помощью **heatmap** и **scatter plot**.

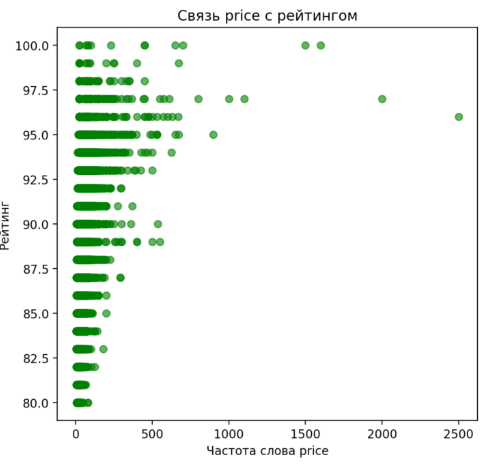




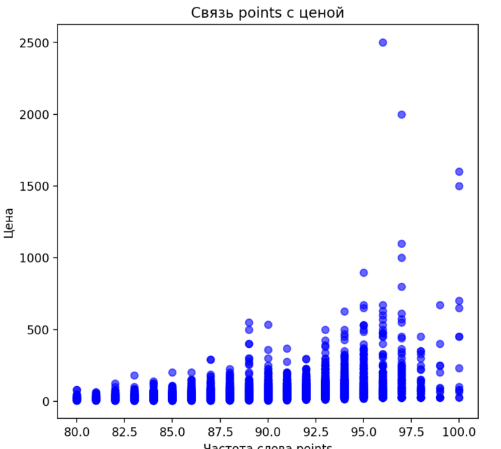
# IV. Анализ корреляций

Были изучены следующие зависимости:

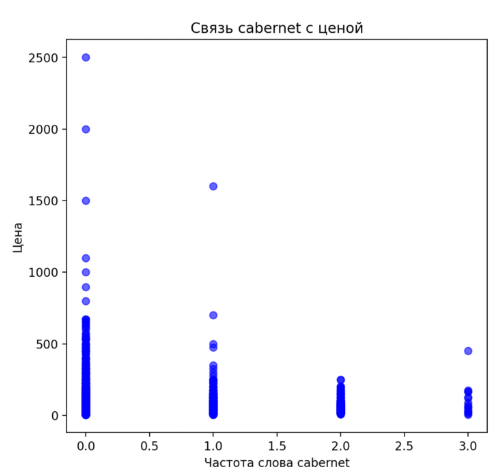
1. **Корреляция между ценой и рейтингом.**



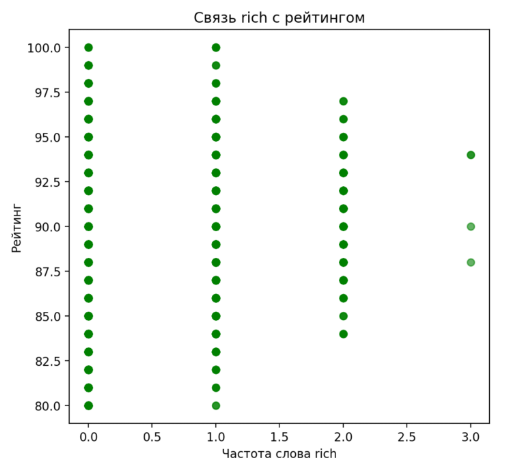
1. **Связь рейтинга с ценой.**



1. **Связь сорта винограда с ценой.**



1. **Зависимость слова с рейтингом.**

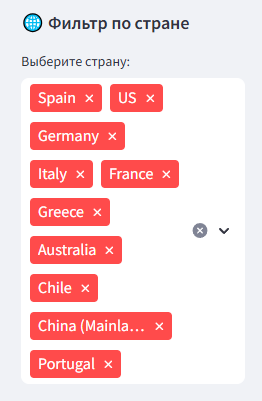


Для наглядности использовались **heatmap** для числовых данных и **scatter plot** для индивидуальных зависимостей.

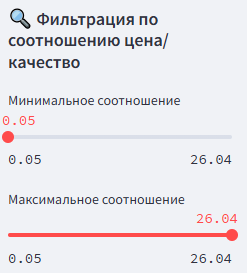
# V. Разработка приложения Streamlit

Разработано интерактивное приложение на **Streamlit** с функционалом:

1. **Фильтры по данным:** 
   * Фильтрация по странам (по умолчанию выбраны 10 основных стран).



* + Фильтрация по соотношению цена/качество.



1. **Генерация визуализаций:** 
   * Гистограмма для баллов.
   * Бар-чарт для средних цен по странам.
   * Составная диаграмма для категорий и регионов.
   * Облако слов для анализа описаний.
2. **Анализ текстовых данных:** 
   * Анализ наиболее частотных слов.
   * Корреляция слов с ценой и рейтингом.

# VI. Результаты и выводы

1. **Выявленные закономерности:**
   * Вина из определенных стран показывают лучшее соотношение цена/качество.
   * Наблюдается положительная корреляция между ценой и рейтингом.
   * Длинные описания часто ассоциируются с более высокими баллами.
2. **Влияние описания на цену и рейтинг:**
   * Слова "фруктовый", "ароматный", "мягкий" чаще встречаются в описаниях вин с высокими баллами и ценами.
3. **Страны и категории вин**:
   * Самые дорогие вина производятся во **Франции** и **Аргентине**.
   * Категории **красные** и **белые вина** занимают доминирующее положение на рынке.

Лабораторная работа продемонстрировала навыки обработки, анализа и визуализации данных. Созданное приложение на **Streamlit** предоставляет удобный инструмент для фильтрации, анализа и визуализации данных о винах, а также помогает выявлять интересные закономерности и зависимости в данных.