САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Институт «Компьютерных наук и кибербезопасности»

«Высшая школа технологий искусственного интеллекта»

Направление «Математическое обеспечение и администрирование информационных систем»

**Отчёт по заданию**

«Введение в науки о данных и визуализации»

Преподаватель: Э.Р. Хольгер

Выполнили:

Бай Борис группа: 5130203/40002 семестр: 2

Мартьянов Иван группа: 5130203/40002 семестр: 2

Тимур Гилязов группа: 5130203/40002 семестр: 2

**Введение:**

Современная наука о данных включает в себя широкий спектр методов и инструментов для анализа, обработки и визуализации данных, что позволяет извлекать полезные знания из больших объемов информации. В рамках данной лабораторной работы будет рассмотрен процесс анализа и визуализации данных, а также применение статистических методов для обработки числовых и категориальных переменных.

Целью лабораторной работы является развитие навыков работы с реальными наборами данных, освоение методов их анализа, визуализации и построения различных статистических моделей. В ходе работы будут использованы инструменты, такие как Jupyter, Python и библиотеки для анализа данных (например, pandas, matplotlib, seaborn), а также другие средства, предназначенные для создания и представления визуальных данных.

Для выполнения лабораторной работы выберем набор данных с числовыми и категориальными признаками, который будет подвергнут различным методам анализа. Важно обратить внимание на обработку пропусков в данных и их визуализацию, что позволит выделить важные закономерности и сделать более информированные выводы.

**Визуализация данных:**

1. **Распределение вин по уровню алкоголя**  
   В данном анализе представлена таблица, которая отображает распределение вин по уровню алкоголя. Эти данные позволяют увидеть, как различается содержание алкоголя в различных образцах вина и насколько это влияет на его качество. Основываясь на визуализации, можно сделать вывод, что большинство вин имеют уровень алкоголя в пределах 9-11%, что соответствует средним данным по красным винам.
2. **Распределение вин по уровню pH**  
   Визуализация распределения вин по уровню pH позволяет понять, какие вина имеют более кислую или щелочную среду. Это важно, так как pH влияет на вкус и сохранность вина. Из диаграммы видно, что большинство вин находятся в диапазоне от 3.0 до 3.5, что соответствует нормальным значениям для красных вин.
3. **Корреляция между химическими характеристиками и качеством вина**  
   В таблице представлена корреляция между различными химическими характеристиками (например, кислотность, сульфаты) и качеством вина. Например, выявлена положительная корреляция между уровнем алкоголя и качеством вина. Это подтверждает, что вина с более высоким содержанием алкоголя часто получают более высокие оценки качества.
4. **Столбчатая диаграмма для категориальных переменных**  
   Столбчатая диаграмма для категориальных переменных, таких как качество вина, помогает наглядно продемонстрировать частоту появления различных оценок качества. Диаграмма показывает, что большинство вин имеет оценку качества в пределах 5-6, что соответствует среднему качеству.
5. **Круговая диаграмма для анализа качества вина**  
   Эта диаграмма показывает процентное распределение вин по различным категориям качества. Как оказалось, большинство вин имеют оценку 5, что может быть связано с тем, что большинство вин на рынке не относятся к премиум-классу, но и не являются полностью низкокачественными.
6. **Диаграмма зависимости качества от уровня алкоголя**  
   На круговой диаграмме видно, как уровень алкоголя влияет на качество вина. Диаграмма подтверждает гипотезу, что вина с более высоким содержанием алкоголя имеют тенденцию к более высоким оценкам качества.
7. **Диаграмма для числовых данных**  
   Диаграмма помогает оценить разбор возрастных данных среди образцов вина. На основе этих данных можно сделать вывод, что виноматериалы с более высоким качеством имеют тенденцию к более высокому содержанию алкоголя и сульфатов.
8. **Гистограмма распределения содержания сульфатов**  
   Гистограмма отображает распределение содержания сульфатов в винах. Преобладает группа с низким содержанием сульфатов, что подтверждается тенденцией к высокому качеству вина при минимальном добавлении сульфатов.
9. **Групповой анализ качества вина по содержанию алкоголя и pH**  
   На данной диаграмме показана зависимость качества вина от сочетания уровня алкоголя и pH. Видно, что вино с высоким содержанием алкоголя и оптимальным pH имеет тенденцию к более высоким оценкам качества.

**Заключение**

В ходе выполнения лабораторной работы был проведен анализ и визуализация данных, связанного с химическими характеристиками красного вина и их влиянием на оценку качества. В этой работе был использован набор данных, содержащий такие параметры, как уровень алкоголя, pH, содержание сульфатов и другие характеристики, которые могут влиять на восприятие и качество вина.

Технически для анализа использовались инструменты Python и библиотеки, такие как pandas для обработки данных, matplotlib и seaborn для визуализации. В первую очередь был выполнен импорт и предварительная обработка данных. Важной частью работы стало выявление и обработка пропущенных значений в наборе данных с использованием методов pandas, таких как .isnull() и .dropna().

Далее был проведен статистический анализ, включая вычисление корреляций между числовыми характеристиками с использованием функции .corr() из pandas. Это позволило определить, какие из химических характеристик наиболее сильно влияют на оценку качества вина. Все корреляции были визуализированы с помощью тепловых карт, что позволило наглядно оценить взаимосвязи между различными переменными.

Для категориальных переменных, таких как качество вина, были построены столбчатые и круговые диаграммы с использованием библиотеки matplotlib. Это дало возможность наглядно представить распределение различных категорий качества и их соотношение с другими переменными, например, с уровнем алкоголя или pH.

Кроме того, для более глубокого анализа были построены гистограммы, показывающие распределение значений различных характеристик, таких как содержание сульфатов и алкоголя, что помогло визуализировать их влияние на качество вина.

В процессе работы использовались следующие основные методы:

1. **Загрузка и предварительная обработка данных** с использованием pandas.
2. **Статистический анализ** данных с помощью функции .corr() для выявления корреляций.
3. **Визуализация данных** с помощью matplotlib и seaborn, включая гистограммы, столбчатые и круговые диаграммы, а также тепловые карты для визуализации корреляций.

Таким образом, техническая реализация анализа и визуализации данных о качестве вина показала важность грамотной обработки данных и их визуального представления. Использование Python и библиотек для анализа данных позволяет эффективно и наглядно выявлять закономерности и делать выводы на основе данных. Этот процесс является неотъемлемой частью научных исследований и может быть адаптирован для дальнейшего анализа данных в различных областях.