МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ государственное БЮДЖЕТНОЕ

образовательное учреждение

высшего образования

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кафедра Автоматизированных систем управления



**Расчетно-графическая работа**

**по дисциплине: «**Методы анализа данных**»**

Выполнили: Проверил:

Студенты: Тягунова В.В. Преподаватель: Ганелина Н.Д.

Ермаков И.В.

Майнгардт А.А.

Группа: АВТ-713

Новосибирск

2020

**Цель работы:** изучение методов факторного анализа.

**Среда выполнения:** SPSS Statistics.

**Задание**

1. Выбрать массив данных (ссылки № 3-4 из списка литературы),

описать параметры.

2. Стандартизировать переменные. Провести корреляционный анализ, сформулировать предположения о наличии скрытых факторов.

3. Провести дескриптивный анализ, оценить близость выборок к нормальной.

4. Изучить пример решения задачи (ссылки №2, 5-10 из списка литературы).

5. Определить факторы методом главных компонент.

6. Изменить факторные нагрузки методом вращения.

7. Определить количество факторов с помощью критерия каменистой осыпи и критерия Кайзера.

8. Интерпретировать результаты.

9. Оформить отчет.

**Описание исходных данных**

В качестве исходных данных был выбран массив, в котором описана информация о качестве вин. Набор данных содержит 1599 наблюдений, данное подтверждение представлено на Рис. 1:

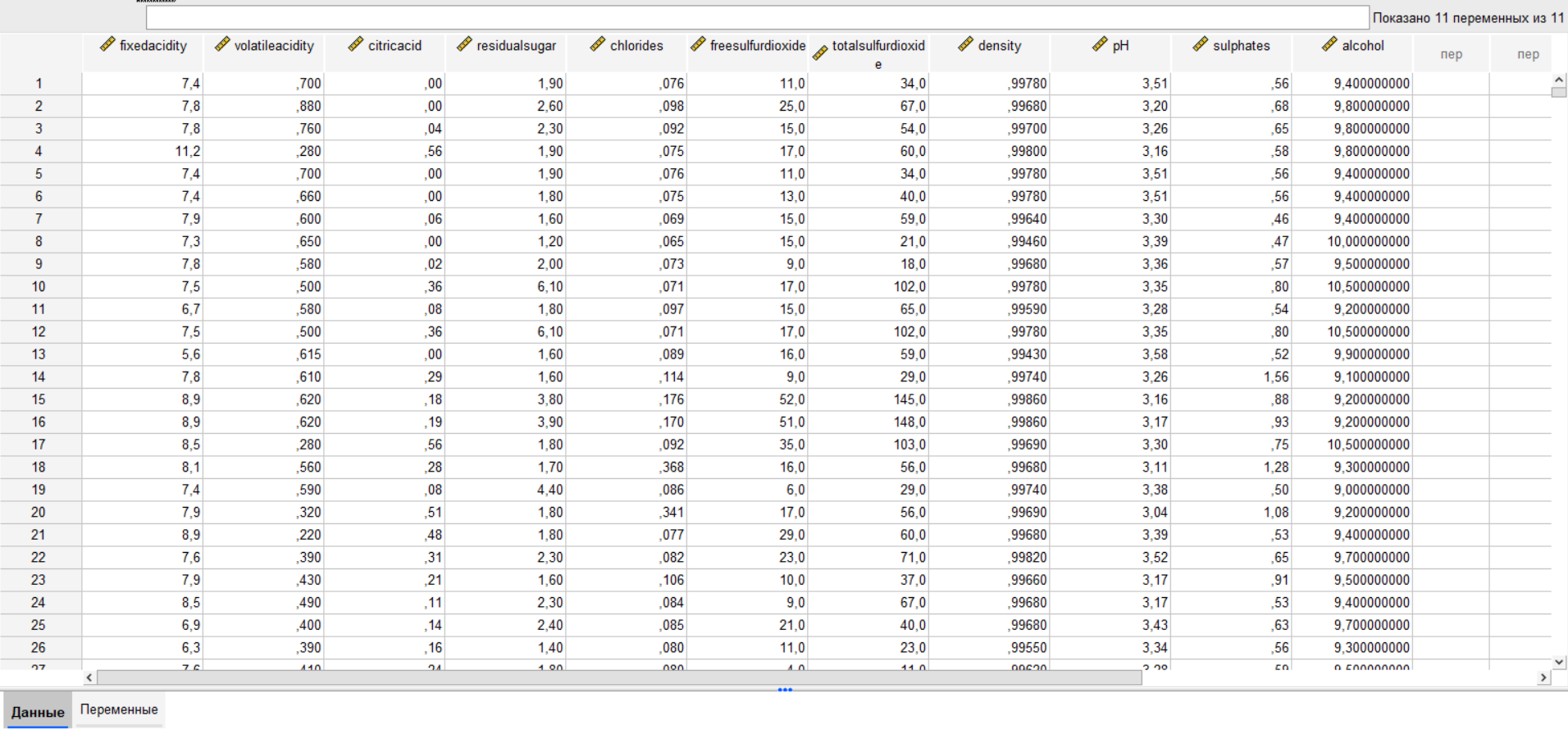


Рис.1. Фрагмент таблицы с исходными данными

В данном массиве присутствует 11 данных:

1 - *фиксированная кислотность (fixed acidity)*

2 - *летучая кислотность (volatile acidity)*

3 - *лимонная кислота (citric acid)*

4 - *остаточный сахар (residual sugar)*

5 – *хлориды (chlorides)*

6 - *свободный диоксид серы (free sulfur dioxide)*

7 - *общий диоксид серы (total sulfur dioxide)*

8 – *плотность (density)*

9 - *мера*[*активности*](https://ru.wikipedia.org/wiki/%D0%90%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D1%8C_(%D1%85%D0%B8%D0%BC%D0%B8%D1%8F))[*ионов*](https://ru.wikipedia.org/wiki/%D0%98%D0%BE%D0%BD)*водорода в*[*растворе*](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D1%82%D0%B2%D0%BE%D1%80)*, количественно выражающая его кислотность (pH)*

10 – *сульфаты (sulphates)*

11 – *алкоголь (alcohol)*

**Результаты дескриптивного анализа**

В основе дескриптивного (описательного) анализа лежат такие статистические показатели как средняя величина, мода, медиана, стандартное отклонение и другие. Они представляют собой характеристики переменных - величин, которые в исследованиях можно измерять, контролировать или варьировать.

Рис.2. Дескриптивный анализ

На рисунке 2 представлена описательная статистика наших данных. Как видно все данные являются ассиметричными (положительная асимметрия, длинный правый хвост). Эксцесс – это мера интенсивности выбросов. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс означает, что выбросы в данных интенсивнее, чем для нормального распределения. Отрицательный эксцесс означает, что выбросы в данных менее интенсивны, чем для нормального распределения. В нашем случае для входящих данных эксцесс больше нуля, это означает, что у всех данных островершинное распределение. Также можно заметить, что для данных плотность и pH будут близки к нормальному распределению.

Из этого можно сделать вывод, что ни одна из выборок не является нормальной. Данный факт может повлиять на правдоподобность наших результатов, так как для факторного анализа используются данные, которые распределены по нормальному закону, в нашем случае это может привести к отклонению модели, также для данных, которые не распределены по нормальному закону, используется многомерное шкалирование, но наши данные более-менее удовлетворяют требованиям факторного анализа, поэтому используем факторный анализ, так как он гораздо более мощная процедура, намного лучше оценивает связи исходных переменных, а также благодаря ему гораздо проще интерпретировать результаты, в многомерном шкалировании очень трудно объяснить, что же значат полученные факторы.

**Корреляционный анализ**

Перед тем, как проводить факторный анализ, можно проверить, существует ли зависимость между переменными: если зависимости не будет обнаружено, факторы выделить не получится.

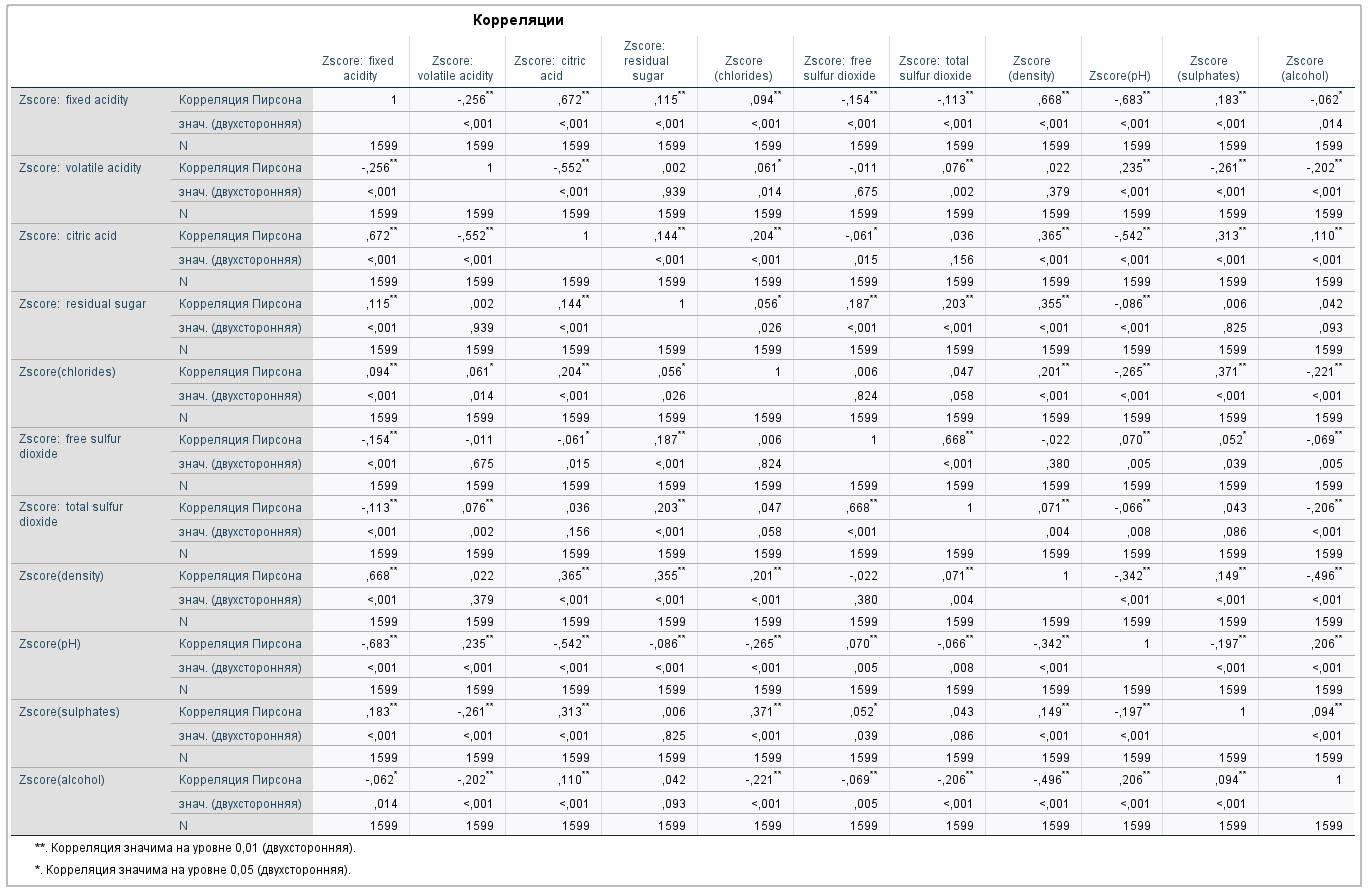


Рис.3. Корреляция

Как видно из рисунка 3, у нас есть умеренно положительная связь между величинами free sulfur dioxide и total sulfur dioxide, их корреляция равна 0,668, а также у величин density и fixed acidity 0,668 и тд. Это означает, что существуют некоторые зависимости.

**Факторный анализ. Матрица факторных нагрузок до и после применения методов вращения.**

Факторный анализ – это процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям, сводится к меньшему количеству независимых влияющих величин, называемых факторами. При этом в один фактор объединяются переменные, сильно коррелирующие между собой. Переменные из разных факторов слабо коррелируют между собой. Таким образом, целью факторного анализа является нахождение таких комплексных факторов, которые как можно более полно объясняют наблюдаемые связи между переменными, имеющимися в наличии.

Две основных цели факторного анализа:

* определение взаимосвязей между переменными;
* сокращение числа переменных необходимых для описания данных.

Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят:

* все признаки должны быть количественными;
* число наблюдений должно быть не менее чем в два раза больше числа переменных;
* выборка должна быть однородна;
* исходные переменные должны быть распределены симметрично;
* факторный анализ осуществляется по коррелирующим переменным.

*Критерий Кайзера*

Согласно критерию Кайзера, отбираются только факторы с собственными значениями равными или большими единицы. Иными словами, фактор дающие дисперсию эквивалентную как минимум дисперсии одной переменной отбираются для дальнейшей работы.



Рис.4. Объясненная совокупная дисперсия

В результате анализа все исходные переменные были разделены на четыре группы (фактора), так как собственные значения именно четырех факторов превосходят 1, также вместе они объясняют около 70,8% дисперсии. По рисунку 4 можно увидеть, что собственные значения четырех факторов превосходят 1, то есть целесообразно выделить именно 4 фактора.

*Критерий каменистой осыпи*

Критерий каменистой осыпи предполагает найти такое место на графике, где убывание собственных значений слева направо замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь».

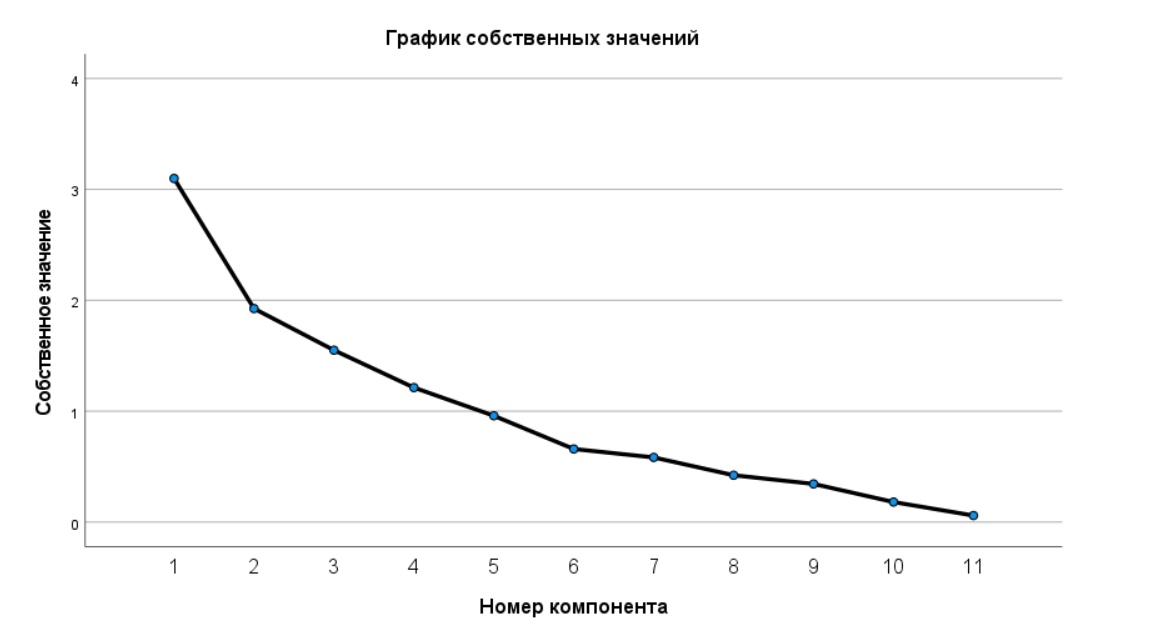


Рис.5. График описывающий критерий каменистой осыпи

График показывает соответствующие собственные значения в системе координат: с 5 по 6 факторы происходит перелом графика. Это подтверждает, что оптимальное количество факторов 5. Но так как покритерию Кайзера первые 4 фактора уже описывают 70% дисперсии, можно оставить 4 фактора.

*Вращение факторов*

Важный результат факторного анализа — матрица факторных нагрузок, также называемая матрицей факторного отображения. Она содержит коэффициенты, используемые для выражения нормированных переменных через факторы. Эти коэффициенты, называемые факторными нагрузками, представляют корреляции между факторами и переменными. Коэффициент с высоким абсолютным значением показывает, что фактор и переменная тесно взаимосвязаны. Коэффициенты матрицы факторных нагрузок можно использовать для интерпретации факторов.

Матрица факторных нагрузок до вращения представлена на рисунке 6.

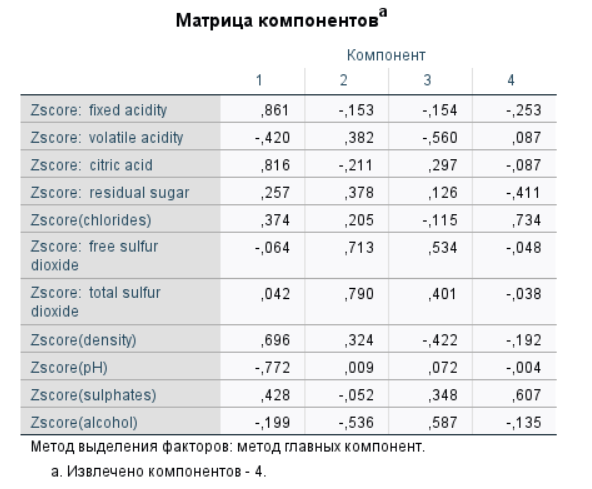


Рис.6. Матрица факторных нагрузок до вращения.

Матрица факторных нагрузок после вращения представлена на рисунке 7.



Рис.7. Матрица факторных нагрузок после вращения



Рис.8. Матрица факторных нагрузок после вращения (с выделенными элементами)

**Интерпретация результатов**

После вращения можно увидеть, что переменные стали коррелироваться хорошо только с одним фактором, до вращения некоторые переменные коррелировали с несколькими факторами.

И теперь распределение по факторам произошло следующим образом:

1) Фиксированная кислотность, лимонная кислота, плотность, мера активности ионов водорода в растворе;

2) Свободный диоксид серы, общий диоксид серы;

3) Летучая кислотность, алкоголь;

4) Хлориды и сульфаты.