Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

|  |
| --- |
| Институт информационных технологий и анализа данных |

наименование института

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 4

по дисциплине:

|  |
| --- |
| **Методы анализа данных** |
| **«Метод главных компонент и факторный анализ»** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил | АСУб-20-2 |  |  |  | Арбакова А.В. |
|  | шифр группы |  | подпись |  | Фамилия И.О. |
| Проверил |  |  |  |  | Бучнев О.С. |
|  | должность |  | подпись |  | Фамилия И.О. |

Иркутск 2022 г

**Цель лабораторной работы:**

1. Выбрать среду программирования для языка Python. Использовать набор данных согласно номеру варианта (табл. 1.1).
2. Загрузить набор данных, отобрать признаки, измеренные в количественной шкале и подготовить данные для анализа. Построить диаграммы рассеяния. С использованием Python выполнить снижение размерности и факторный анализ данных. При этом необходимо:

* выполнить стандартизацию данных;
* построить корреляционную матрицу и график «тепловая карта», определить коллинеарные признаки;
* определить, какую часть общей дисперсии описывают главные компоненты, при пороговом значении 80 % общей дисперсии определить число главных компонент;
* найти две главные компоненты, построить диаграммы рассеяния для новой системы координат; сравнить диаграммы рассеяния с исходными диаграммами;
* выполнить факторный анализ без вращения, вывести матрицу нагрузок и матрицу общностей; выделить факторы и отнести к ним признаки, дать интерпретацию факторам;
* выполнить факторный анализ с вращением для выбранного числа факторов, вывести матрицу нагрузок, матрицу общностей; сравнить результат факторного анализа с результатом, полученным без вращения факторов.

1. Выполнить анализ полученных на каждом этапе результатов и оформить отчет по лабораторной работе.

**Язык программирования:** Python

**Номер варианта:** 3

|  |  |  |
| --- | --- | --- |
| **№  варианта** | **Название набора данных** | **Число  факторов** |
| 3 | Boston.csv | N = 4 |

# **Ход выполнения работы**

Снижение размерности – это процесс уменьшения анализируемого множества данных до размера, оптимального с точки зрения решаемой задачи и используемой аналитической модели.

Сокращение размерности может потребоваться когда данные избыточны в информационном плане, т.е. задачу можно решить с тем же уровнем эффективности и точности, но используя меньший объем данных. Это позволяет урезать время и вычислительные затраты на решение задачи. Другой случай связан со слишком большими вычислительными затратами, требуемыми для обработки множества данного размера. Эта ситуация типична для алгоритмов, вычислительная сложность которых экспоненциально растет с увеличением числа наблюдений (т.е. немасштабируемых). Если в первом случае достаточно просто отобрать из всего множества столько признаков и записей, сколько надо, то во втором нужно сократить исходное множество до такого объема, который обеспечил бы реализуемость его обработки невзирая на потерю полезной информации.

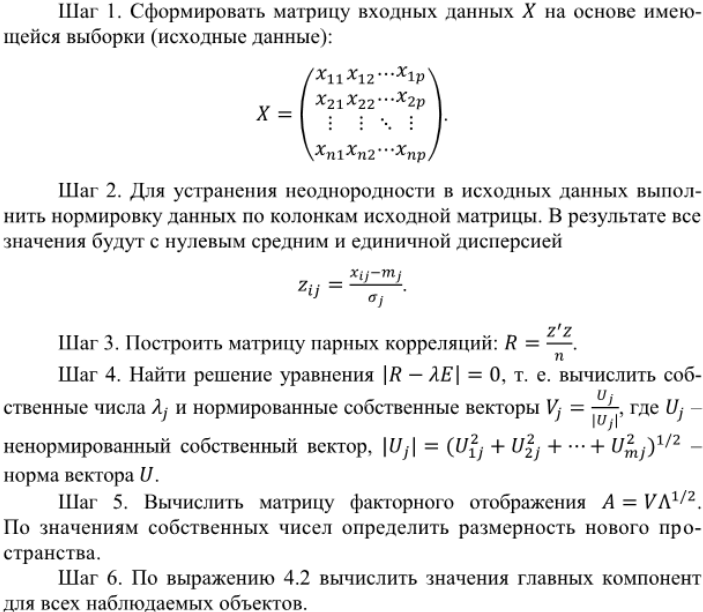
Причины для снижения размерности:

* необходимость наглядного представления (визуализации) исходных данных, что достигается их проецированием на специально подобранное трехмерное пространство (К = 3), плоскость (К =2) или прямую;
* стремлением к лаконизму исследуемых моделей, обусловленному необходимостью упрощения счета и интерпретации полученных статистических выводов;
* необходимостью существенного сжатия объемов хранимой статистической информации (без существенных потерь в се информативности).

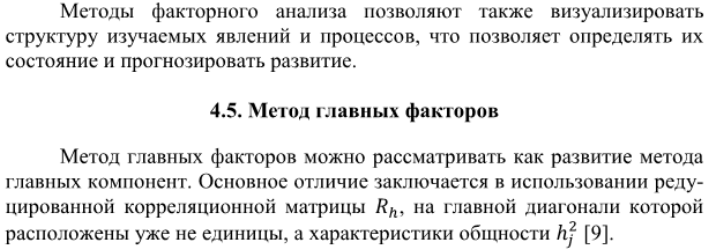
Один из способов решения задачи снижения размерности – нахождение главных компонент. Суть задачи выделения главных компонент заключается в том, чтобы среди всех линейных комбинаций р признаков наблюдаемых объектов выделить гораздо меньшее число К таких, изменчивость которых в значительной степени описывает изменчивость всего первоначального набора признаков в целом. В дальнейшем можно использовать эти найденные комбинации (которые и называются главными компонентами), для классификации или решения других задач.

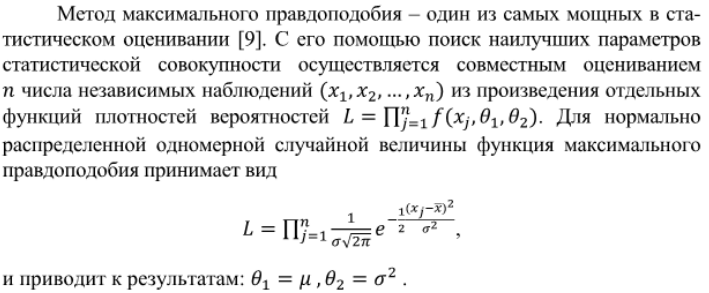
Математическая модель метода главных компонент базируется на логичном допущении, что значения множества взаимосвязанных признаков порождают некоторый общий результат.

Алгоритм метода главных компонент:



Факторный анализ занимается определением относительно небольшого числа скрытых факторов, изменчивостью которых объясняется изменчивость всех наблюдаемых показателей каждого объекта. В этом случае оцениваемые факторы можно считать причинами, а наблюдаемые признаки – следствиями. Результаты факторного анализа принимают успешными в том случае, когда большое число признаков удалось объяснить малым количеством причин. Таким образом, факторный анализ, как и метод главных компонент, направлен на снижение размерности.





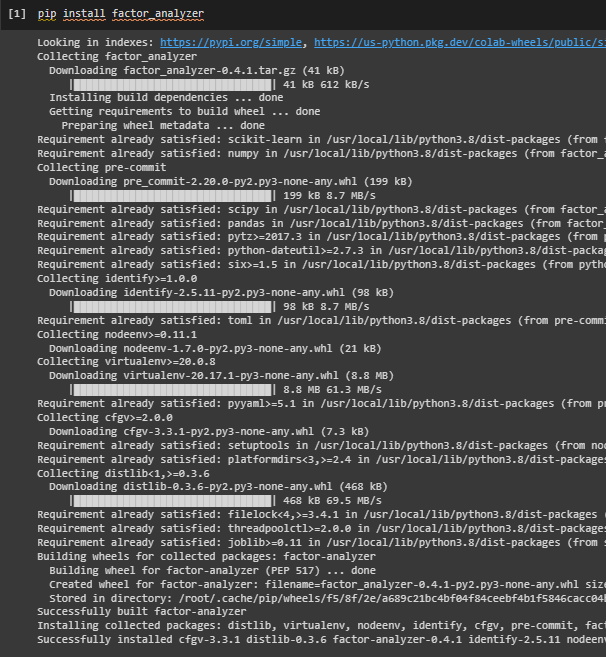
Факторные нагрузки – корреляции переменных с выделенным фактором. Чем больше величина факторной нагрузки, тем сильнее связь этой переменной с фактором, то есть факторная нагрузка (степень взаимосвязи) показывает, насколько данная переменная обусловлена действием соответствующего фактора.

Общность – это стандартизированная внешняя нагрузка элемента. По аналогии с r-квадратом Пирсона, возведенная в квадратную нагрузку представляет собой процент отклонения в этой индикаторной переменной, объясняемой этим фактором. Чтобы получить процент дисперсии всех чисел, учитываемых факторов, сложите сумму возведенных в квадрат факторных нагрузок для этого фактора (столбец) и разделите на количество факторов.

Вращение факторов. Целью вращения является получение простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору и малое по всем остальным факторам. Нагрузка отражает связь между переменной и фактором, являясь подобием коэффициента корреляции. Значение нагрузки лежит в пределах (-1; 1).

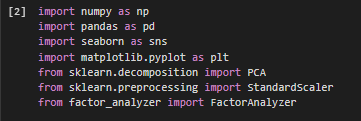
Варимакс (Varimax) — это такое ортогональное вращение, при котором происходит минимизация количества переменных с высокой факторной нагрузкой.

Выполним установку пакета factor\_analyser:



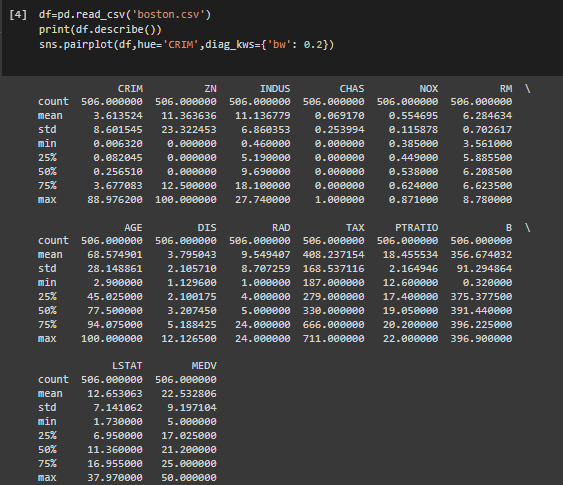
Изображение 1 – Установка пакета factor\_analyser:

Подключим необходимые библиотеки и модули:

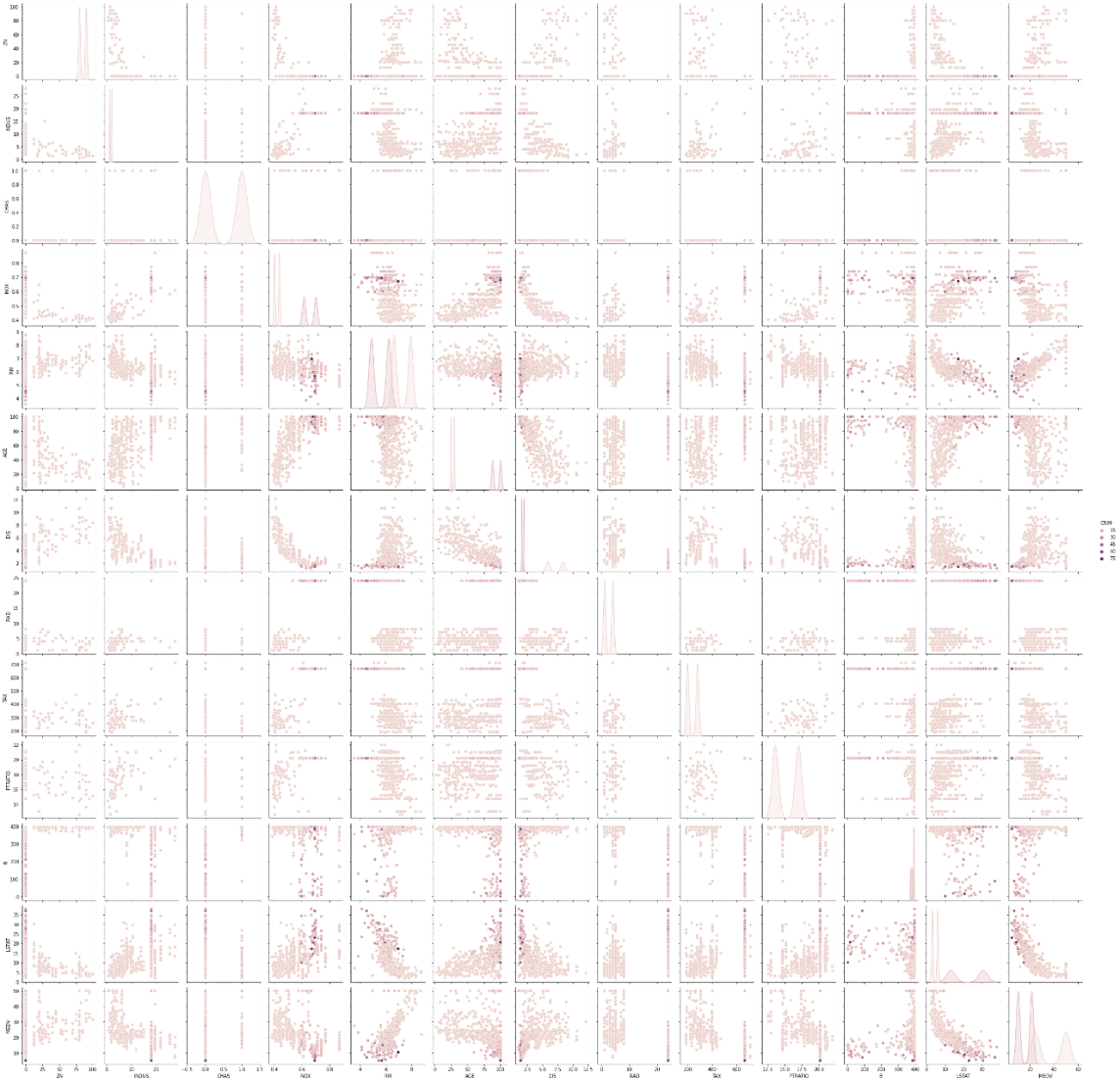


Изображение 2 – Подключение библиотек и модулей

Загрузим данные, подготовим их для анализа. Получим значения основных показателей описательной статистики и построим диаграммы рассеяния для признаков, измеренных в количественной шкале:

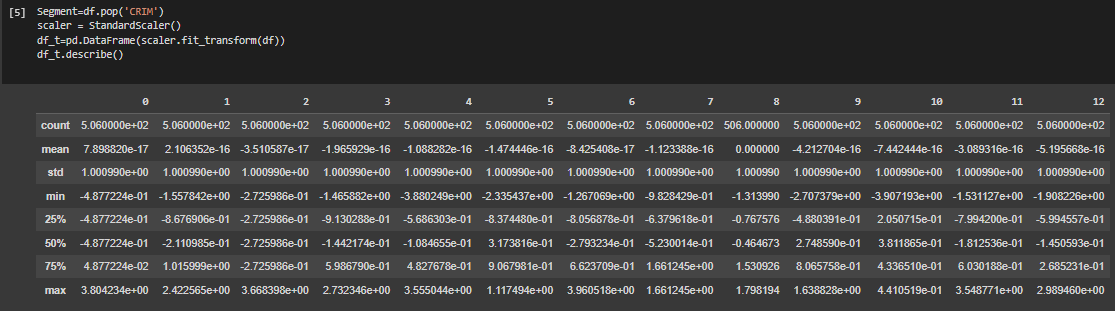


Изображение 3 – Основных показатели описательной статистики



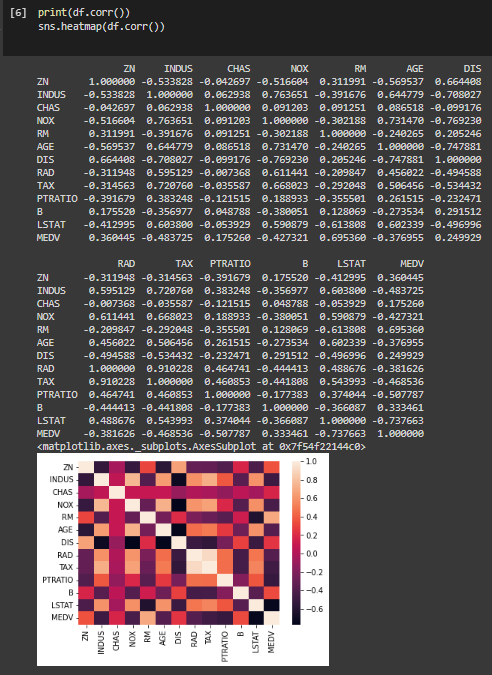
Изображение 4 – Диаграммы рассеяния для признаков, измеренных в количественной шкале

Отделим переменную отклика от набора данных и выполним его стандартизацию:



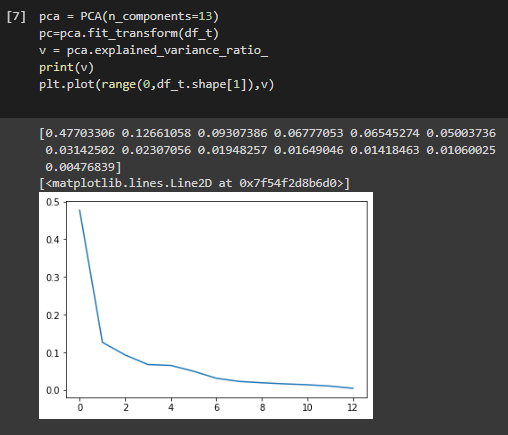
Изображение 5 – Отделение переменной отклика от набора данных и выполнение ее стандартизации

Построим корреляционную матрицу и график «тепловая карта»:



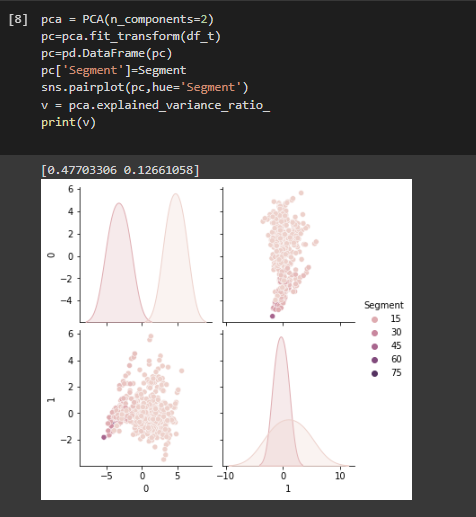
Изображение 6 – Корреляционная матрица и тепловая карта

Определим, какую часть общей дисперсии описывают главные компоненты:



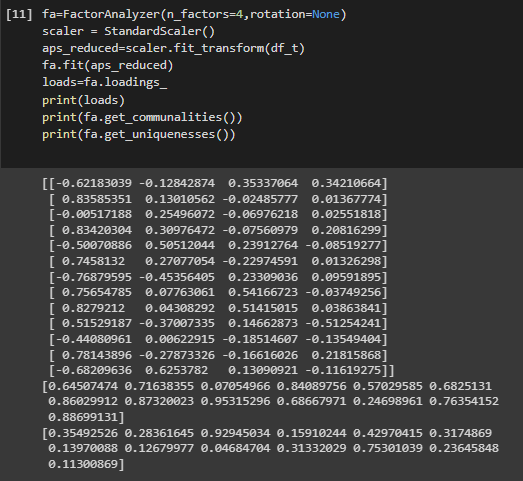
Изображение 7 – Часть общей дисперсии, которую описывают главные компоненты

Для того, чтобы «забрать» 80 % общей дисперсии, нужны 4 компонент. Найдем две главные компоненты, построим диаграммы рассеяния для новой системы координат:



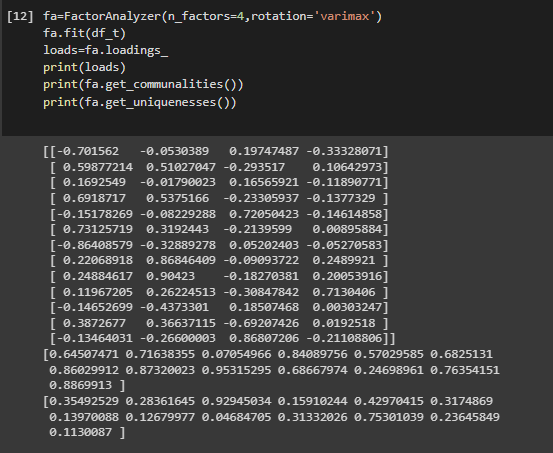
Изображение 8 – Диаграммы рассеяния для новой системы координат

Выполним факторный анализ для четырех факторов без вращения, выведем матрицу нагрузок и матрицу общностей:



Изображение 9 – Матрица нагрузок и матрица общностей, факторный анализ без вращения

Выполним факторный анализ с вращением varimax для четырех факторов, выведем матрицу нагрузок и матрицу общностей:



Изображение 10 – Матрица нагрузок и матрица общностей, факторный анализ с вращением

# **Заключение**

В ходе выполнения лабораторной работы по теме «Метод главных компонент и факторный анализ» была выполнена цель работы реализации метода главных компонент и факторный анализ.

Было изучены такие темы как: проблема размерности, преимущества снижения размерности, главные компоненты, математическая модель, алгоритм метода главных компонент, назначение факторного анализа, метод главных факторов, метод максимального правдоподобия, факторные нагрузки, общности и вращение факторного пространства.

Знания были закреплены во время использования среды разработки Google Colab и использования языка программирования Python, в котором было выполнено задание лабораторной работы.

# **Список литературы**

1. Бучнев О. С. Методы анализа данных: лабораторный практикум – Иркутск, 2022. – 115 с
2. <https://colab.research.google.com>
3. С.Э. Мастицкий, В.К. Шитиков Статистический анализ и визуализация данных с помощью языка R, Хайдельберг – Лондон – Тольятти, 2014 г. – 401 с