Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

**Иркутский национальный исследовательский   
технический университет**

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

|  |
| --- |
| Отчет по дисциплине |
| «Методы анализа данных» |
| по теме: |
| «Факторный анализ и главные компоненты» |
|  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил студент группы |  | АСУб-20-1 |  |  |  | Устюжанин В. М. |
|  |  | Шифр группы |  | Подпись |  | И.О. Фамилия |
| Проверил преподаватель |  |  |  |  |  | Осипова Е. А. |
|  |  |  |  | Подпись |  | И.О. Фамилия |

Иркутск 2022 г.

**Содержание**

[Постановка задачи 3](#_Toc118469870)

[Подготовка данных для анализа 4](#_Toc118469871)

[Вывод 12](#_Toc118469872)

# Постановка задачи

А) Выбрать среду программирования для языка Python. Использовать набор данных согласно номеру варианта (таблица 1).

Б) 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.

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

Варианту №7 принадлежит следующий датасет:

|  |  |  |
| --- | --- | --- |
| 7 | Bodyfat.csv | N = 5 |

Density – Плотность тела (в г/см3);

Bodyfat – Процент жира в организме;

Age – Возраст;

Weight – Вес;

Height – Рост;

Neck – охват шеи (в см.);

Chest – охват груди (в см.);

Abdomen – охват талии (в см.);

Hip – охват ягодиц (в см.);

Thigh – охват бедер (в см.);

Knee – охват колен (в см.);

Ankle – охват лодыжки (в см.);

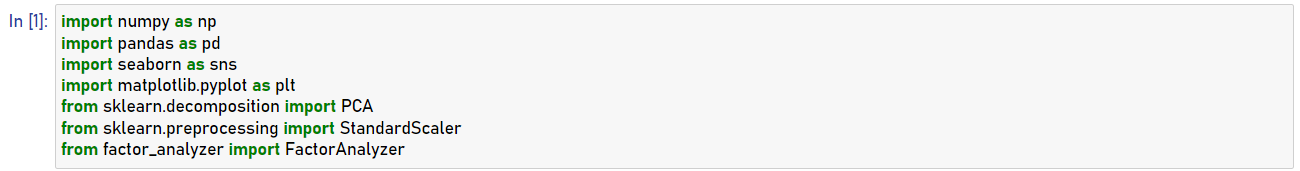
Biceps – охват бицепса (в см.);

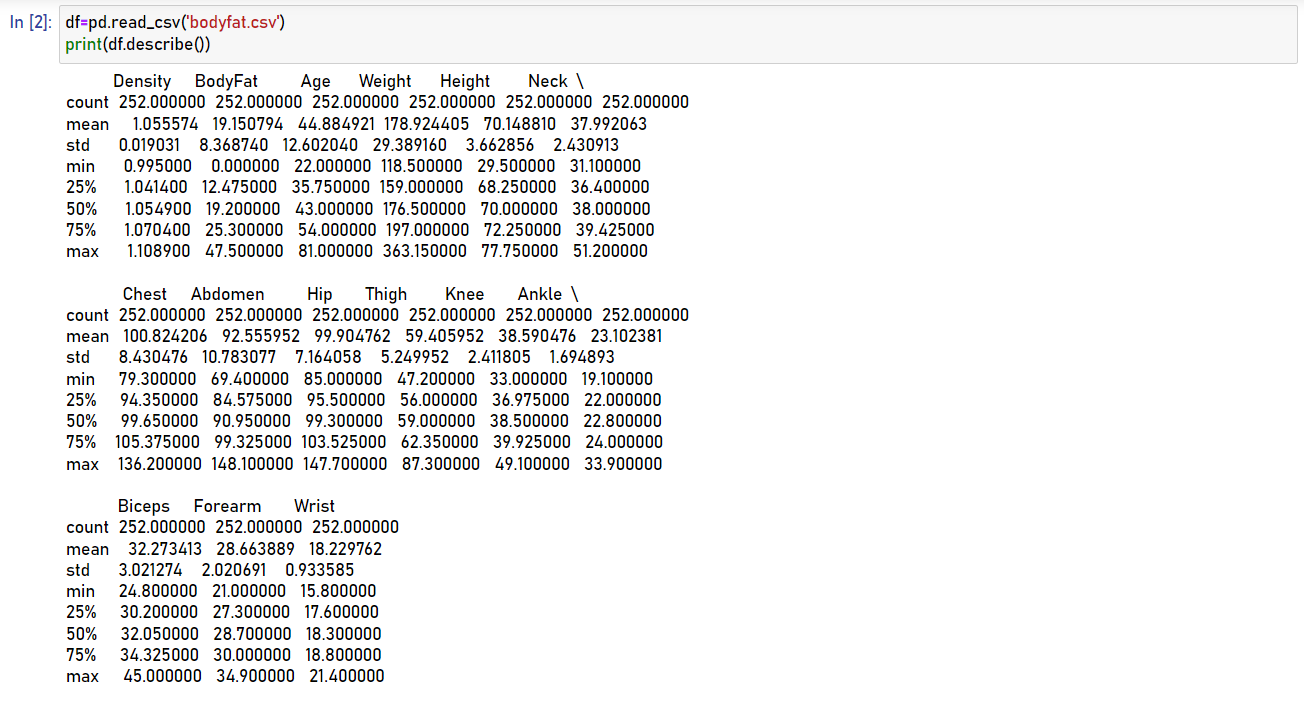
Forearm – охват предплечья (в см.);

Wrist – охват запястий (в см.);

# Выполнение лабораторной работы

Подключим библиотеки, необходимые для выполнения задания.

  
Рисунок 1 – Подключение библиотек

  
Рисунок 2 - Значения основных показателей описательной статистики

Построим диаграмму рассеивания (фрагмент):

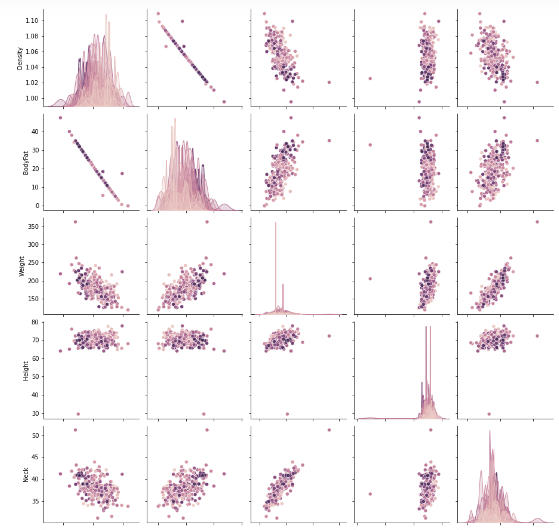


Рисунок 3 – Диаграмма рассеивания

Отделим переменную отклика и выполним стандартизацию набора данных:

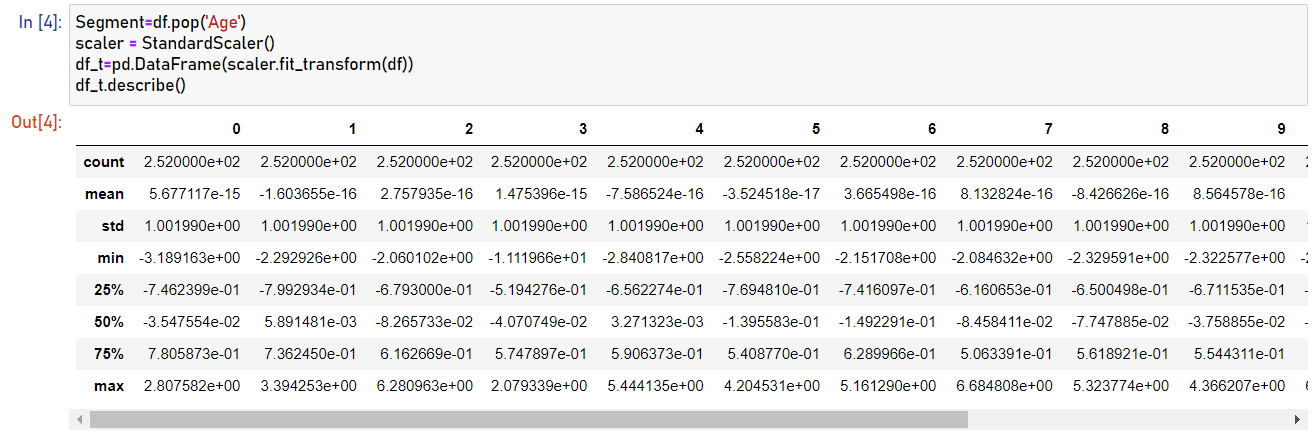


Рисунок 4 – Нормализация (фрагмент)

Построим корреляционную матрицу:

****

Рисунок 6 – Корреляционная матрица

График «тепловая карта»:

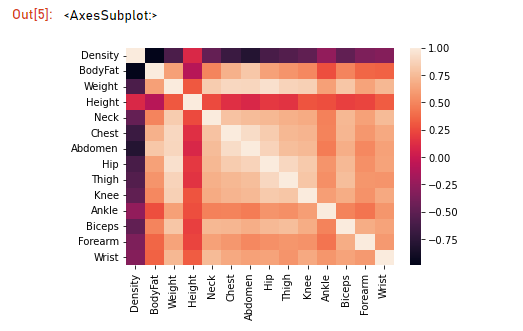


Рисунок 7 – Тепловая карта

Определим, какую часть общей дисперсии описывают главные компоненты:

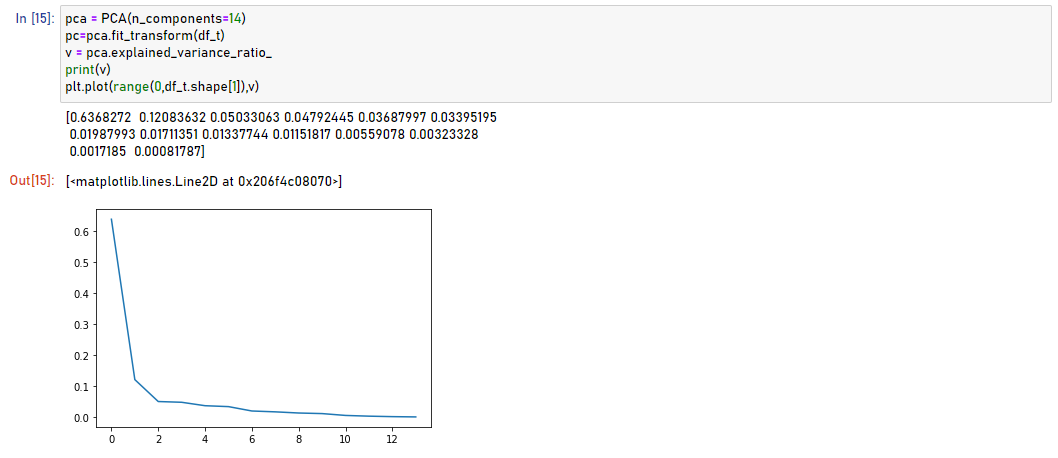


Рисунок 8 - Общая дисперсия для главных компонент

Вычислим минимальное число компонент чтобы забрать минимум 80% общей дисперсии:

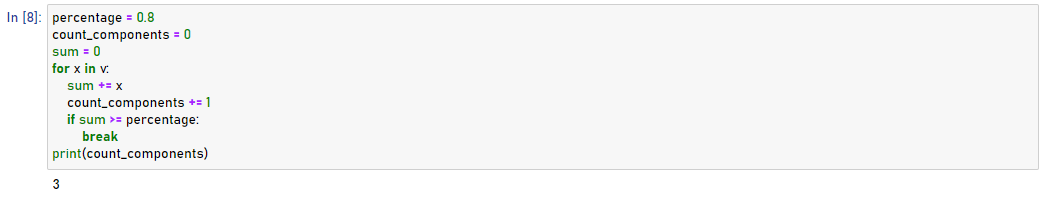


Рисунок 9 - Число компонент для 80% дисперсий

Найдем 3 главных компонента, построим диаграммы рассеяния для новой системы координат:

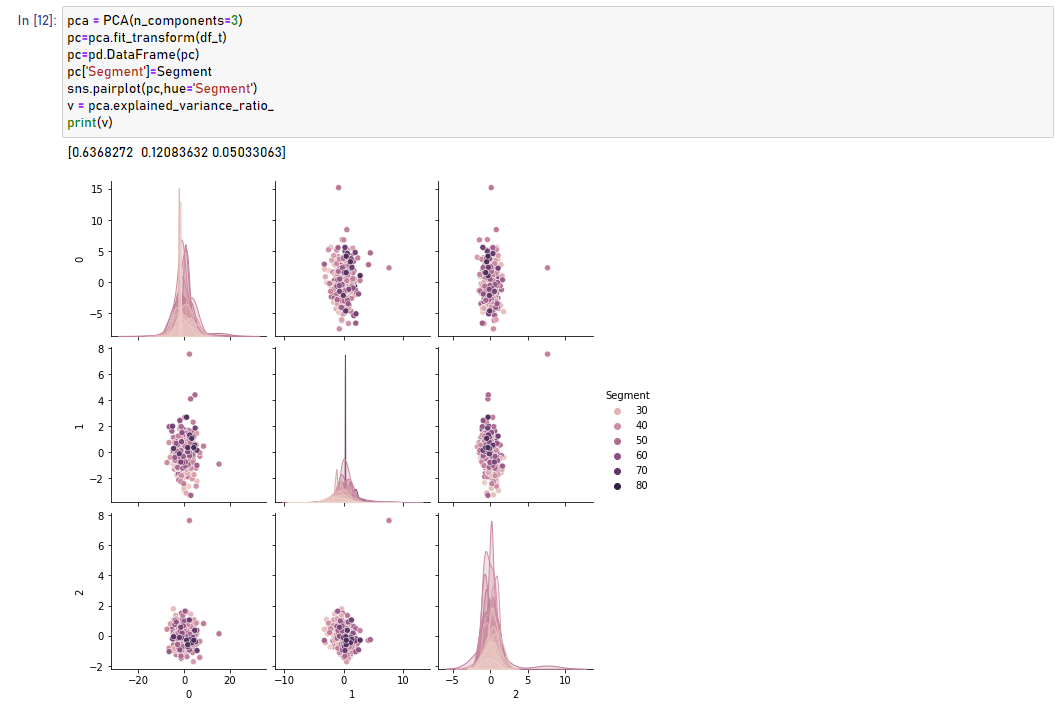


Рисунок 10 - Диаграмма рассеяния для новой системы координат

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.

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

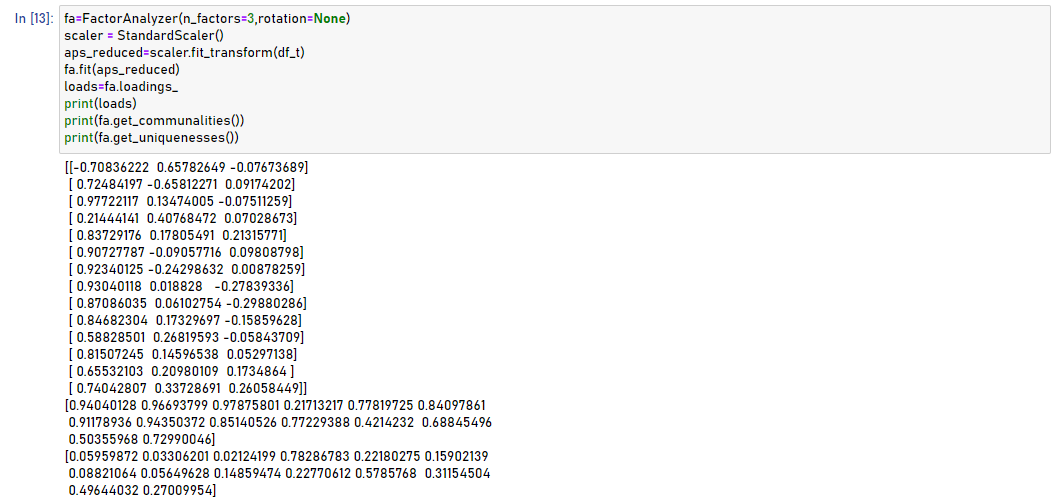


Рисунок 11 - Матрицы нагрузок и общностей без вращения

Проинтерпретируем факторные нагрузки и общности.

1-ый фактор сильнее всего зависит от следующих причин: Возраст, рост, охват шеи, охват груди, охват талии, охват ягодиц, охват бедер, охват колен, охват лодыжки и охват бицепса.

2-ой фактор: не обнаружено сильных зависимостей.

3-ой фактор: не обнаружено сильных зависимостей.

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



Рисунок 12 - Матрицы нагрузок и общностей с вращением «varimax»

Проинтерпретируем факторные нагрузки и общности.

1-ый фактор стал сильнее всего зависеть от охвата запястий.

2-ой фактор стал сильнее всего зависеть от процента жира в организме.

3-ой фактор стал сильнее всего зависеть от охвата талии и охвата ягодиц.

Сравнивая результаты с предыдущими, видим, что есть небольшие изменения.

# Вывод

В ходе работы был выполнен метод главных компонент и факторный анализ. С помощью метода главных компонент была снижена размерность данных. С помощью факторного анализа выявлены 3 новых фактора и их основные зависимости от существующих.