Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Институт информационные системы и анализ данных

Центр программной инженерии

**ОТЧЁТ**

к лабораторной работе №4 по дисциплине:

|  |
| --- |
| «Методы анализа данных» |
| Факторный анализ и главные компоненты |
| наименование темы |

Вариант 13

Выполнилстудент ИСТб-21-1 Левин А.

номер группы подпись И. О. Фамилия

дата

Проверил Доцент Е.А. Осипова

Должность подпись И. О. Фамилия

дата

Иркутск 2023

**Оглавление**

[Задание 3](#_Toc135850799)

[1 Ход работы 4](#_Toc135850800)

[1.1 Исходные данные 4](#_Toc135850801)

[Заключение. 13](#_Toc135850802)

[Список использованных источников 14](#_Toc135850803)

# Задание

А. Выбрать среду программирования для языка Python. Использовать набор данных согласно номеру варианта (табл. 1.1).

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.

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

|  |  |  |
| --- | --- | --- |
| 5 | Penguins.csv | N=2 |

# 1 Ход работы

# Исходные данные

Получение исходных данных для варианта 13. Hitters.csv.

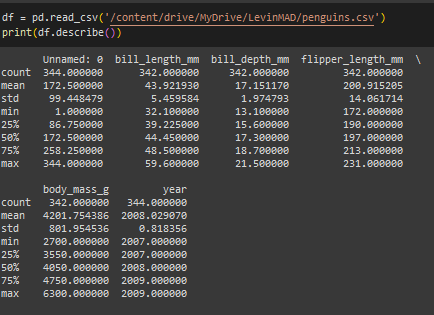


Рисунок 1 – Исходных данные.

Диаграмма рассеяния по дата сету Hitters.csv не может быть построена с учетом нынешних мощностей компьютера и доступных мощностей серверов Google collab

Стандартизация данных и вывод квантилей.

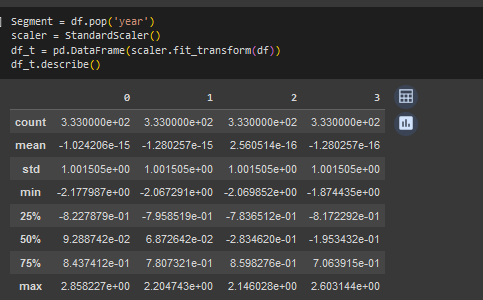


Рисунок 2 – Стандартизация данных.

Строим матрицу корреляции и выводим график «тепловая карта»

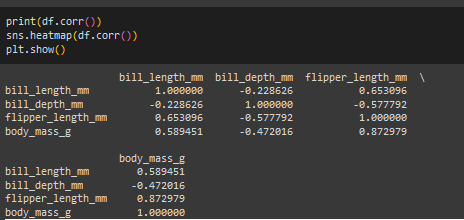


Рисунок 3 – Матрица корреляции.

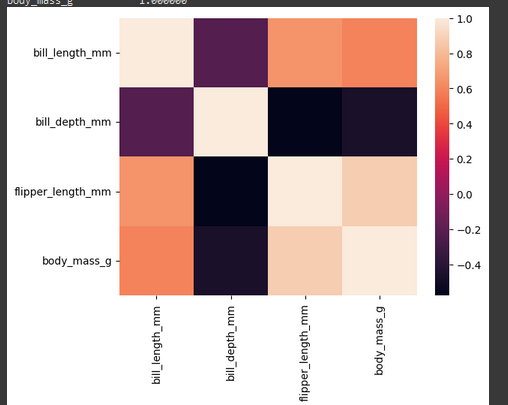


Рисунок 4 – График «Тепловая карта».

Найдем количество главных компонент:

Этот код на Python вычисляет количество главных компонент, необходимых для описания 80% общей дисперсии в данных с использованием метода кумулятивной суммы собственных значений.

Краткое описание кода:

* **cumsum\_var** - это массив, содержащий накопленную сумму собственных значений после их упорядочивания по убыванию.
* **np.argmax(cumsum\_var >= 0.8)** возвращает индекс первого элемента, удовлетворяющего условию **cumsum\_var >= 0.8**.
* Прибавление 1 к результату **np.argmax(cumsum\_var >= 0.8)** дает количество компонент, необходимых для описания 80% общей дисперсии в данных, так как индексы в Python начинаются с нуля.

Таким образом, **n\_components** будет содержать количество главных компонент, необходимых для описания 80% общей дисперсии в данных.

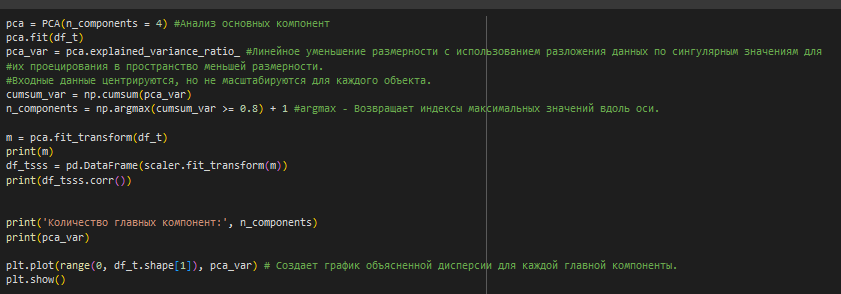


Рисунок 5 – Поиск количества главных компонент

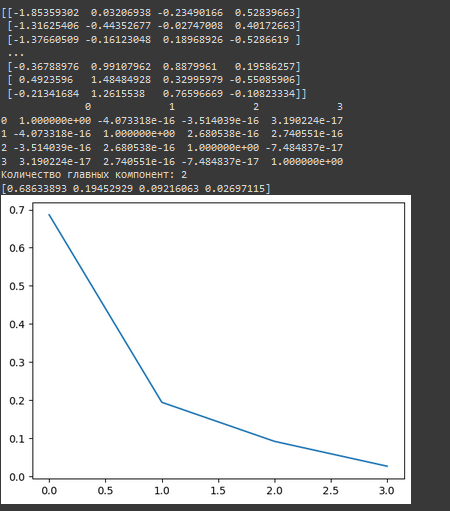


Рисунок 6 – Массив значений, которые показывают, какую часть дисперсии описывают главные компоненты, а также график объясненной дисперсии.

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

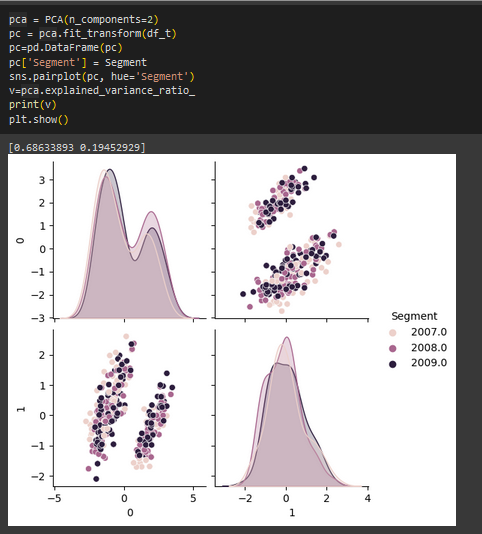


Рисунок 7 – Массив главных компонент, а также диаграмма разброса этих компонент.

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

Матрица нагрузок (без вращения) показывает, как каждый признак связан с каждым фактором, при условии, что факторы не поворачивались. Каждый признак представлен строкой, а каждый фактор представлен столбцом. Значение в каждой ячейке матрицы показывает, насколько данный признак связан с данным фактором.

Значения в матрице нагрузок определяются корреляцией между каждым признаком и каждым фактором, поэтому значения, близкие к 1 или -1, указывают на сильную связь между признаком и фактором, а значения, близкие к 0, указывают на отсутствие связи.

Матрица общностей представляет собой диагональную матрицу, в которой на главной диагонали расположены элементы, характеризующие долю дисперсии каждого из исходных признаков, объясненную соответствующей главной компонентой.

В данном случае, результат показывает, что первая главная компонента объясняет 85% дисперсии исходных данных, вторая компонента - 48.93%, седьмая - 95%, восьмая - 94%, а десятая - 96%, одиннадцатая – 96%. Это говорит о том, что седьмая, восьмая, десятая и одиннадцатая главные компоненты могут описать большую часть исходных данных, а остальные компоненты объясняют менее значимую часть дисперсии исходных данных.

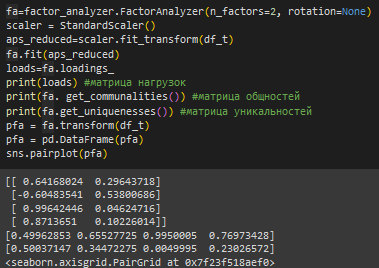


Рисунок 8 – Матрица нагрузок без вращения.

После проведения факторного анализа строим диаграмму разброса для факторного анализа без вращения

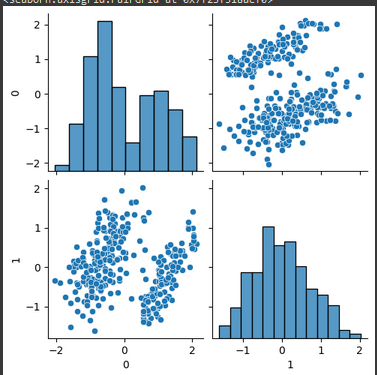


Рисунок 9. Диаграмма разброса факторного анализа без вращения

Вращение в факторном анализе является важным этапом, который позволяет упростить и улучшить интерпретацию результатов исследования, повышая качество и точность выводов.

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

Матрица нагрузок (с вращением) показывает связь между исходными переменными и выделенными факторами. Каждый элемент матрицы представляет собой коэффициент, который указывает на взаимосвязь между соответствующей исходной переменной и конкретным фактором.

В данном случае, матрица нагрузок (с вращением) имеет размерность 16x2, где 16 - количество исходных переменных, а 2 - количество выделенных факторов. Каждый столбец матрицы соответствует одному фактору.

Также важно отметить, что вращение факторов изменяет матрицу нагрузок, но не влияет на объясняемую дисперсию.

Интерпретация матрицы нагрузок (с вращением) зависит от конкретной задачи и контекста. Например 1 фактор сильно связан с 4, 6, 7,8, 9, 10, 11, 12, 16 переменными.

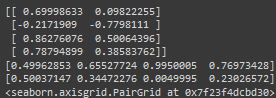


Рисунок 10 – Матрица нагрузок с вращением «varimax».

Матрица общностей в факторном анализе с вращением показывает, какую часть дисперсии каждая переменная объясняет своими факторами. Значения на диагонали матрицы общности соответствуют общей дисперсии каждой переменной, а значения вне диагонали - объясненной дисперсии.

В данном случае матрица общностей (с вращением) имеет те же значения, что и матрица общностей без вращения, что говорит о том, что вращение не привнесло дополнительной информации и не изменило долю объясненной дисперсии каждой переменной.



Рисунок 11 – Матрица общностей с вращением «varimax».

После проведения факторного анализа строим диаграмму разброса для факторного анализа с вращением.

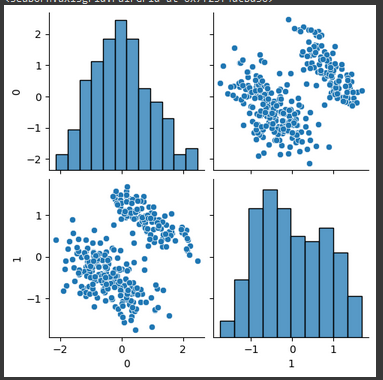


Рисунок 12. Диаграмма разброса факторного анализа с вращением

# Заключение.

В ходе выполнения лабораторной работы был произведен анализ набора данных с использованием языка программирования Python. Были отобраны признаки, измеренные в количественной шкале, проведена их стандартизация, а также построены диаграммы рассеяния.

Для дальнейшего анализа было выполнено снижение размерности и факторный анализ данных. Была построена корреляционная матрица и график "тепловой карты", что позволило определить коллинеарные признаки.

Было определено, что 4 главные компоненты описывают около 80% общей дисперсии. Построены диаграммы рассеяния для новой системы координат и сравнены с исходными диаграммами.

Также был выполнен факторный анализ без вращения, выведены матрица нагрузок и матрица общностей, выделены факторы и произведена интерпретация результатов.

Наконец, был выполнен факторный анализ с вращением для выбранного числа факторов, выведены матрица нагрузок и матрица общностей, а также сравнен результат с факторным анализом без вращения.

Анализ результатов позволил получить более детальное представление о данных и выделить наиболее значимые факторы, описывающие переменные. В целом, выполнение лабораторной работы позволило более глубоко понять принципы работы методов снижения размерности и факторного анализа данных.

# Список использованных источников

1 Петров А.В., Бучнев О.С. Лабораторный практикум — Иркутск: ФГБОУ ВО ИРНИТУ, 2022. — 114 с.