МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Национальный исследовательский ядерный университет «МИФИ»

(НИЯУ МИФИ)

Институт Финансовых Технологий и Экономической Безопасности

Кафедра Финансового Мониторинга

Лабораторная работа №3:

По курсу «Макростатический анализ и прогнозирование»

Работу выполнил: студент группы С18-712: Гуляев Е.А.

Проверила: Домашова Д. В.

Москва 2021

# 1. Постановка задачи

Субъект РФ характеризуются следующими признаками:

X1 - соотношение мужчин и женщин (оценка на конец года; на 1000 мужчин приходится женщин);

Х2 – численность населения с денежными доходами ниже величины прожиточного минимума в процентах от общей численности населения субъекта Российской Федерации;

Х3 – общие коэффициенты брачности на 1000 человек населения

Х4 – среднедушевые денежные доходы населения (в месяц; рублей);

Х5 – коэффициенты миграционного прироста на 10 000 человек населения;

Х6 – зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1000 человек);

Х7 – зарегистрировано преступлений средней тяжести (на 1000 человек);

Х8 – зарегистрировано преступлений особой тяжести (на 1000 человек);

Х9 – предварительно расследовано преступлений, совершенных в состоянии алкогольного опьянения (на 1000 человек);

Х10 - зарегистрировано преступлений экономической направленности (на 1000 человек).

Ставится задача на основании статистических данных по показателям, соответствующим нужному варианту, снизить размерность признакового пространства методом главных компонент, обеспечив уровень информативности новой системы признаков не ниже 70%.

# 2. Ход работы.

Первым делом стандартизируем исходные данные, чтобы привести их к одному масштабу.

Если исходные признаки, по которым производится классификация объектов, имеют разные единицы измерения, то необходимо перейти к стандартизованным переменным одним из следующих способов:

; ; ; ; ; ,

где  - исходное значение j-го признака на *i*-ом объекте наблюдения;

 - нормированное значение исходного *j*-го признака на *i*-ом объекте наблюдения;

 - среднее значение *j*-го признака;

 - выборочное среднеквадратическое отклонение *j*-го признака;

 - максимальное значение *j*-го признака;

 - минимальное значение *j*-го признака.

Была рассчитана матрица корреляций для исходных данных. Матрицы представлены на Рисунках 1.1 и 1.2.



Рисунок 1.1 – Результат расчета корреляционной матрицы (Statistica).

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 1.2 – Результат расчета корреляционной матрицы (Python).

Теперь проверим гипотезу о незначимости матрицы корреляций. Для этого посчитаем собственные числа для данной матрицы. Результаты представлены на Рисунках 2.1, 2.2.



Рисунок 2.1 – Результаты расчета оценок собственных чисел корреляционной матрицы (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 3.1 – Результаты расчета оценок собственных чисел корреляционной матрицы (Python)

Затем были посчитаны наблюдаемое и критическое значение статистики Хи-квадрат, в результате гипотеза о незначимости корреляционной матрицы была отвергнута. Результаты представлены на Рисунках 3.1 и 3.2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Hi2набл | Hi2крит1 | Hi2крит2 | alpha | k | n | det(R) = |
| 335,427656 | 65,64662 | 118,1359 | 0,05 | 10 | 85 | 0,01577 |

Рисунок 3.1– Проверка гипотезы о незначимости корреляционной матрицы в Excel

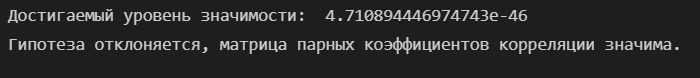


Рисунок 3.2– Проверка гипотезы о незначимости корреляционной матрицы в Python

Дальше были посчитаны доверительные интервалы для собственных значений корреляционной матрицы, результаты представлены на рисунках 4.1 и 4.2.

|  |  |  |
| --- | --- | --- |
| Левая граница | Собственные числа | Правая граница |
| 2,111283219 | 2,749797 | 3,941959377 |
| 1,630294209 | 2,123342 | 3,043908787 |
| 1,068937369 | 1,392215 | 1,99580409 |
| 0,939566873 | 1,223719 | 1,754257512 |
| 0,630045196 | 0,820589 | 1,176352156 |
| 0,478192663 | 0,622812 | 0,892829552 |
| 0,353980689 | 0,461035 | 0,660914406 |
| 0,239425795 | 0,311835 | 0,447029916 |
| 0,121819404 | 0,158661 | 0,227447998 |
| 0,104416249 | 0,135995 | 0,194954711 |

Рисунок 4.1 – Доверительные интервалы собственных чисел, посчитанные в Excel

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 4.2 – Доверительные интервалы собственных чисел, посчитанные в Python

Второе собственное значение попадает в доверительный интервал первого, поэтому проверим гипотезу о кратности собственных чисел (Рисунок 4.3 - 4.4).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Проверка гипотезы о кратности сз | | | | |
| N | r | df | z | p-value |
| 85 | 2 | 2 | 297,8389 | 0 |

Рисунок 4.3 – Гипотеза о кратности собственных чисел (Excel)

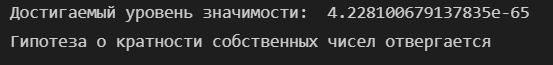


Рисунок 4.4 – Гипотеза о кратности собственных чисел (Python)

Посчитаем информативность (Рисунок 5.1 и 5.2).

|  |  |
| --- | --- |
| Оценка информативности | |
| I\_1(z(x)) = | 0,274979651 |
| I\_2(z(x)) = | 0,487313898 |
| I\_3(z(x)) = | 0,626535402 |
| I\_4(z(x)) = | 0,748907317 |

Рисунок 5.1 – Оценка информативности (Excel)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 5.2 – Оценка информативности (Python)

Информативность превышает 70% при использовании 4 главных компонент. Проверим гипотезу о достаточности 3 главных компонент с помощью критерия Кайзера (Рисунок 6.1 и 6.2).

|  |  |
| --- | --- |
| Критерий Кайзера = | 4 |

Рисунок 6.1 – Критерий Кайзера (Excel)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 6.2 – Критерий Кайзера (Python)

Построим график «осыпи». Результаты представлены на рисунках 7.1 и 7.2.

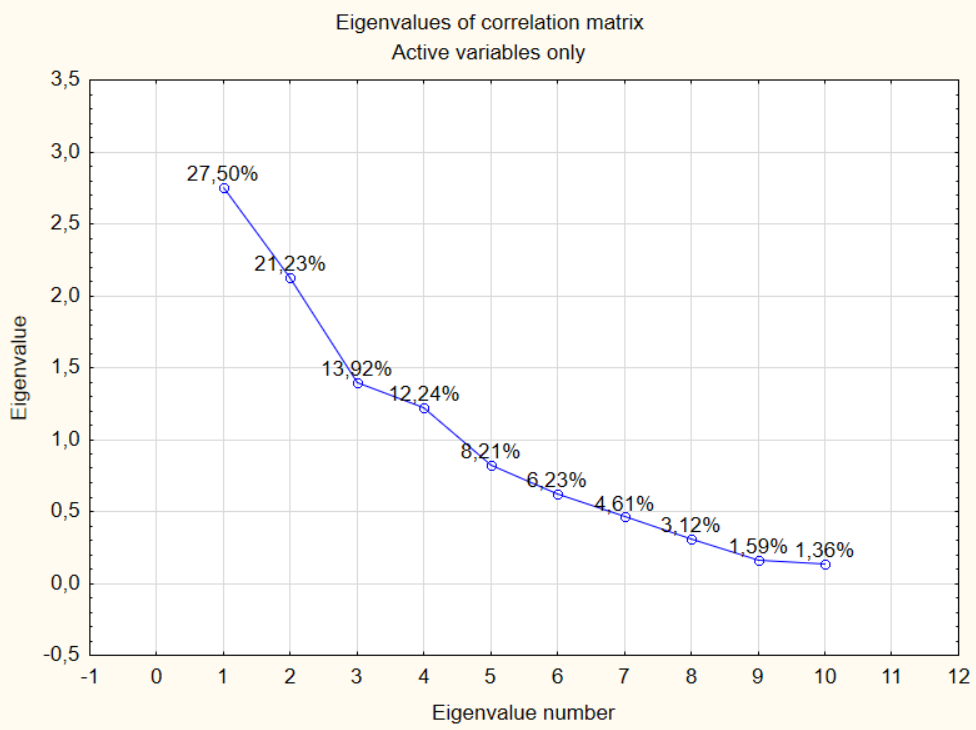


Рисунок 7.2 – График «осыпи» (Statistica)

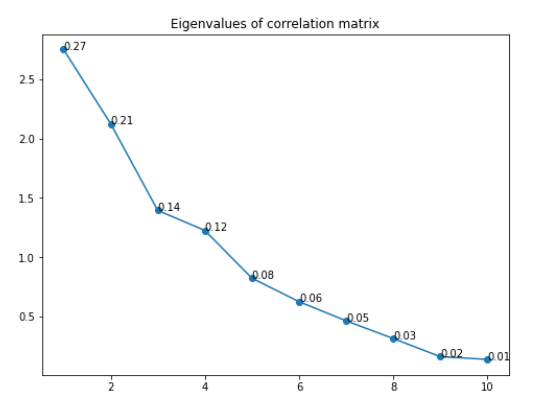


Рисунок 7.2 – График «осыпи» (Python)

Определим вклад каждой главной компоненты в суммарную дисперсию исходных признаков (Рисунок 8.1 и 8.2).



Рисунок 8.1 – Вклады главных компонент в суммарную дисперсию исходных признаков (Statictica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 8.2 – Вклады главных компонент в суммарную дисперсию исходных признаков (Python)

Рассчитаем собственные вектора корреляционной матрицы. Результаты представлены на рисунке 9.



Рисунок 9 – Результаты расчета собственных векторов корреляционной матрицы (Statistica)

После этого посчитаем коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Рисунок 10).



Рисунок 10.1 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 10.2 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Python)

Рассчитаем матрицу нагрузок . Результаты представлены на рисунках 11.1 и 11.2.



Рисунок 11.1 – Результаты расчета элементов матрицы нагрузок (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 11.2 – Результаты расчета элементов матрицы нагрузок (Python)

X1 - соотношение мужчин и женщин (оценка на конец года; на 1000 мужчин приходится женщин);

Х2 – численность населения с денежными доходами ниже величины прожиточного минимума в процентах от общей численности населения субъекта Российской Федерации;

Х3 – общие коэффициенты брачности на 1000 человек населения

Х4 – среднедушевые денежные доходы населения (в месяц; рублей);

Х5 – коэффициенты миграционного прироста на 10 000 человек населения;

Х6 – зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1000 человек);

Х7 – зарегистрировано преступлений средней тяжести (на 1000 человек);

Х8 – зарегистрировано преступлений особой тяжести (на 1000 человек);

Х9 – предварительно расследовано преступлений, совершенных в состоянии алкогольного опьянения (на 1000 человек);

Х10 - зарегистрировано преступлений экономической направленности (на 1000 человек).

Проведем анализ матрицы нагрузок. Так как расчеты проводятся на основании корреляционной матрицы, то элементы матрицы нагрузок являются коэффициентами корреляции исходных признаков и главных компонент. Как видно из таблицы, между исходными признаками и последними 7 главными компонентами не наблюдается тесной связи (не имеется значений >0,7). Поэтому выделим первые 3 главные компоненты.

Первая главная компонента тесно отрицательно связана (модуль коэффициента корреляции больше 0,7) с четырьмя исходными признаками: x6 (Число преступлений, связанных с оборотом наркотиков), x7 (Число преступлений средней тяжести), x8 (Число преступлений особой тяжести), x9 (Число преступлений, совершенных в состоянии алкогольного опьянения). Поэтому первую главную компоненту можно интерпретировать как «Преступность».

Вторая главная компонента тесно отрицательно связана с признаком x2 (Доля населения с доходами ниже прожиточного минимума) и отрицательно с признаком х4 (Среднедушевые доходы населения), поэтому вторую главную компоненту можно интерпретировать «Доходы».

Третья главная компонента тесно положительно с признаком х10 (Число экономических преступлений), поэтому третью главную компоненту можно интерпретировать «Экономическая преступность»

Рассчитаем матрицу индивидуальных значений центрировано - нормированных главных компонент. Результаты представлены на рисунках 12.1 и 12.2.

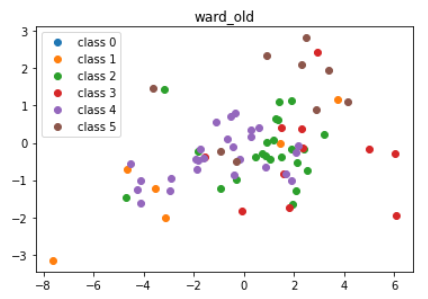


Рисунок 12.1 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 12.2 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Python)



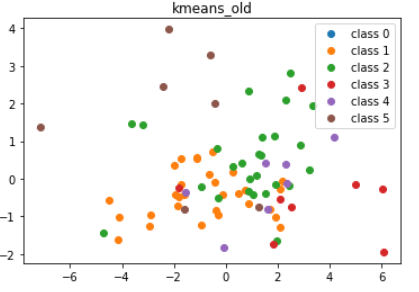
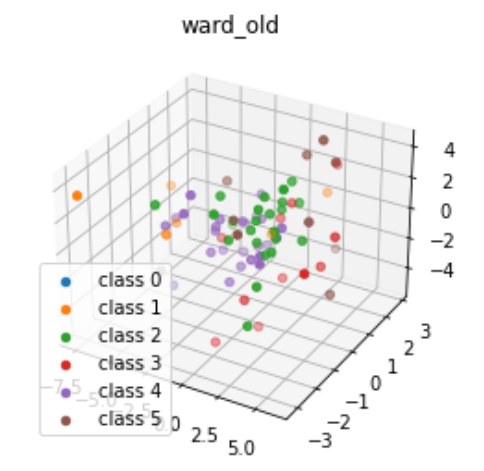


Рисунок 13.1, 13.2 – точечные графики старых кластеров в новой системе координат (2 главных компоненты)





Рисунок 14.1, 14.2 – точечные графики новых кластеров в новой системе координат (2 главных компоненты)



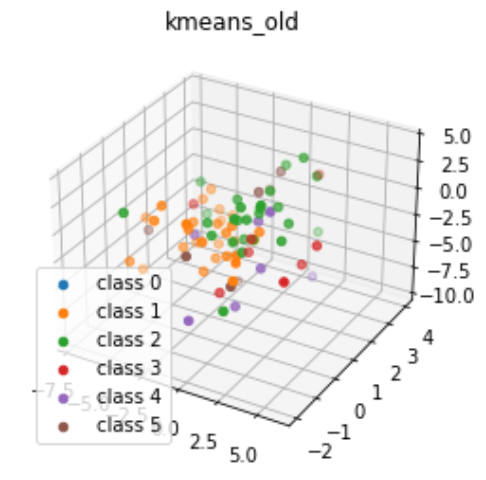
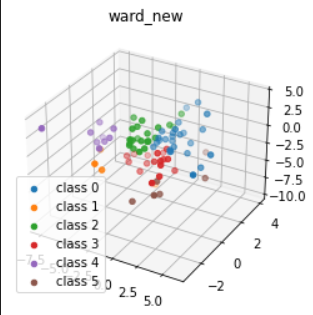


Рисунок 15.1, 15.2 – точечные графики старых кластеров в новой системе координат (3 главных компоненты)



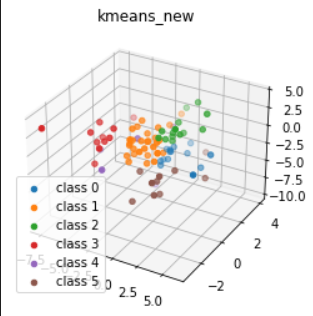


Рисунок 16.1, 16.2 – точечные графики новых кластеров в новой системе координат (3 главных компоненты)

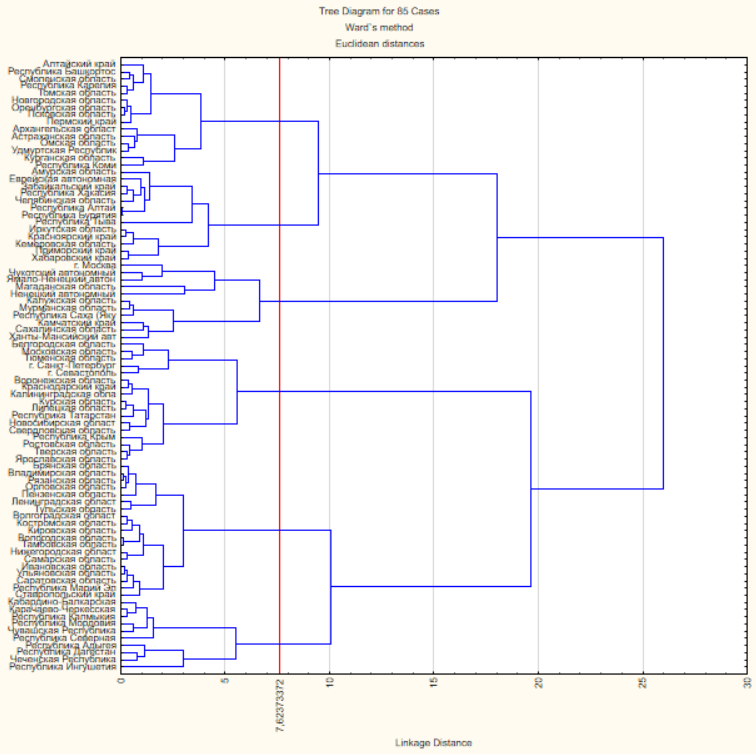


Рисунок 17 – Дендрограмма для кластеризации методом Уорда, полученная по методу главных компонент