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 №4 по курсу**

**«Макростатический анализ и прогнозирование»**

|  |  |
| --- | --- |
| **Выполнил студент группы С21-702:** | Нашивочников А.В. |
| **Проверил:** | Домашова Д.В. |

**Оглавление**

[Задание к лабораторной работе 3](#_Toc87857400)

[Порядок выполнения лабораторной работы 3](#_Toc87857401)

[Матрица нагрузок 9](#_Toc87857402)

[Вывод 25](#_Toc87857403)

## **Задание к лабораторной работе**

Субъекты РФ характеризуются социально-экономическими показателями, обозначение и наименование которых приведены в таблице А.1. Значения показателей для 84 субъектов приведены в таблице А.2. 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%.

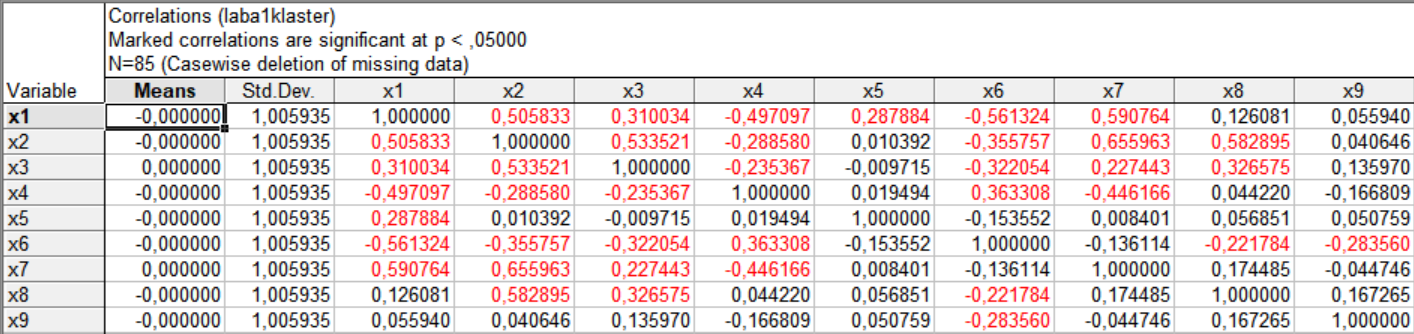
## **Порядок выполнения лабораторной работы**

Порядок выполнения лабораторной работы включает следующие показатели для анализа:

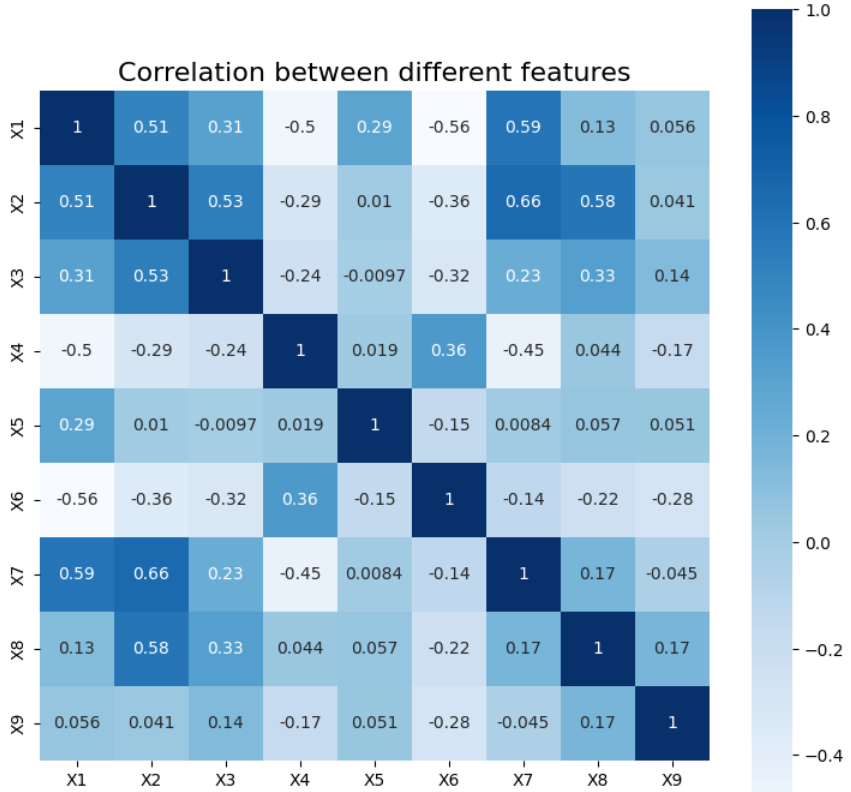
1. X1 – брачность (число браков на 1000 человек)
2. X2 - средний размер назначенных пенсий
3. X3 – заболеваемость на 1000 человек населения
4. X4 - процент безработных мужчин
5. X5 - число посещений музея на 1000 человек населения
6. X6 - число населения на одну больничную койку
7. X7 – потребительские расходы в среднем на душу населения тыс. рублей
8. X8 – количество врачей на 1000 человек населения
9. X9 – число спортивных сооружений на 1000 человек

Поскольку исходные признаки отличаются масштабом измерения, то будем рассматривать вектор центрировано-нормированных признаков x\* = (x1\*, x2\*,…, xk\*)T  и на основе исходной матрицы данных X рассчитаем оценку корреляционной матрицы.

Таблица 1 – Результаты расчета корреляционной матрицы





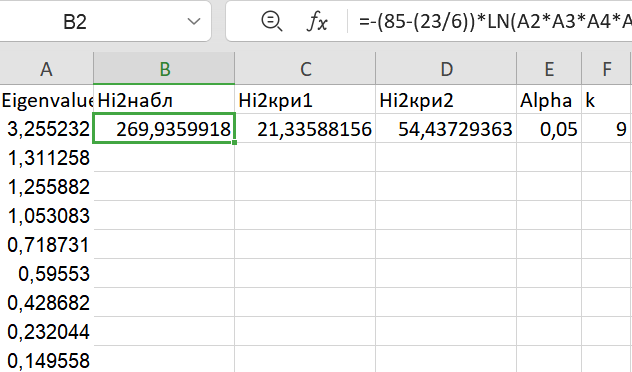


Далее согласно алгоритму, предполагая, что выборка извлечена из нормально распределенной генеральной совокупности, на уровне значимости α = 0,05 проверим гипотезу о незначимости корреляционной матрицы. **H0: Rx = Е;**

**H1: Rx ≠ E.**

Для проверки гипотезы потребуются оценки собственных чисел корреляционной матрицы.

Расчет наблюдаемого значения проводится в Excel на основании полученных оценок собственных значений, которые приводятся на рисунке 1.



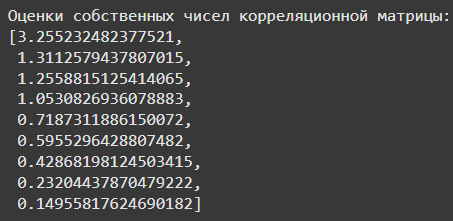


Рисунок 1 – Проверка гипотезы о незначимости корреляционной матрицы

Наблюдаемое значение рассчитывается по формуле:

χ2 = -(n - (2k+5))ln||,

где || - определитель матрицы , равный произведению оценок собственных чисел матрицы;

k – число факторов;

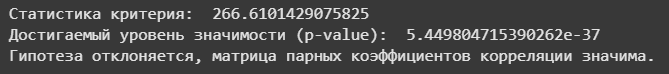
n – объем выборки.

Наблюдаемое значение составило χ 2набл =269,94. Критические значения χкр1 и 2 χкр2 определяются из уравнений:



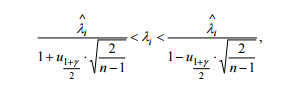
Для решения этих уравнений необходимо воспользоваться функцией ХИ2ОБР(вероятность, ) пакета Excel. Вероятность рассчитывается как 100\*(1-)% для χкр1 и 100\*()% для χкр2 . Число степеней свободы ν = . Критические точки принимают следующие значения:

Так как , то гипотеза  отвергается, матрица парных коэффициентов корреляции значима.



С вероятностью γ = 0,95 построим доверительные интервалы для собственных чисел.

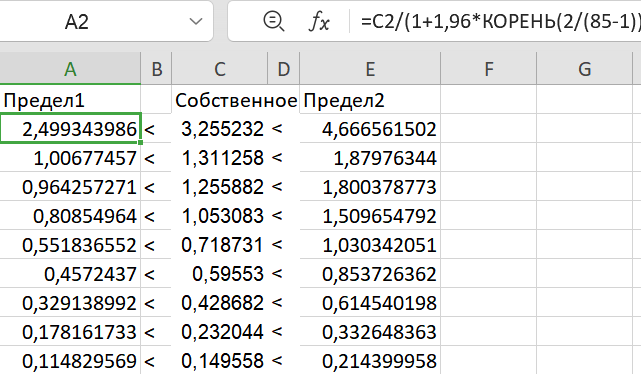
Доверительный интервал для i-ого собственного числа λi при большом объеме выборки имеет вид:

,

где – квантиль уровня стандартного нормального распределения;

n – объем выборки.

Квантиль уровня q можно найти с помощью функции НОРМСТОБР(q) пакета Excel. Получаем: u0,975 = 1,96. Доверительные интервалы для собственных чисел имеют вид, представленный на рисунке 2.



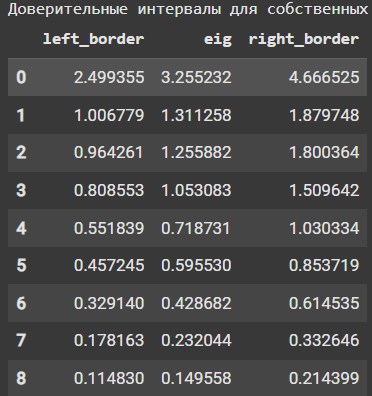
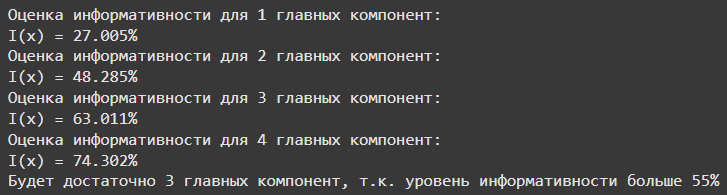


Рисунок 2 – Доверительные интервалы собственных чисел

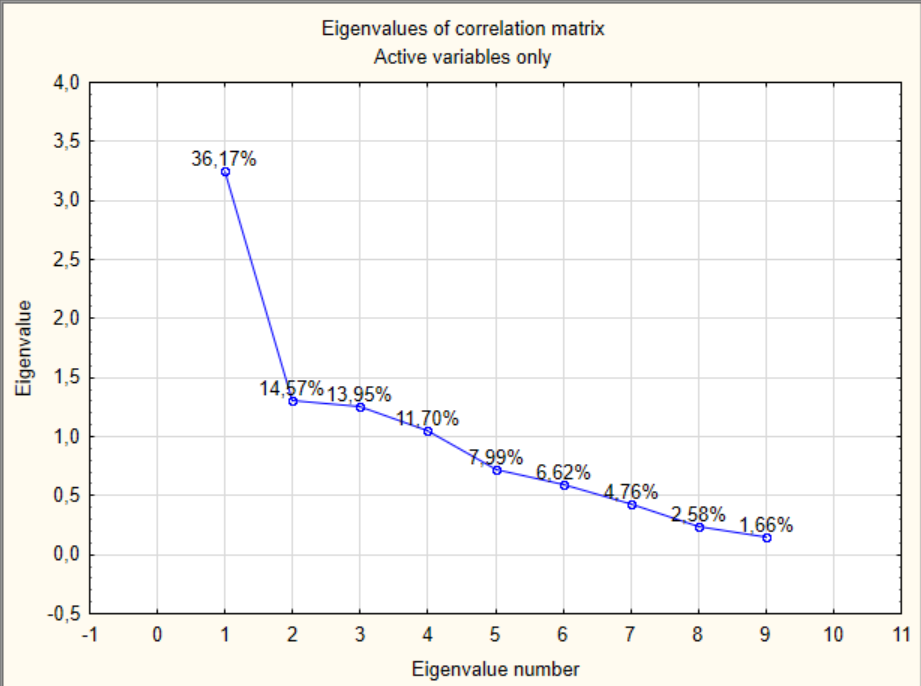
Третье собственное число попадает в доверительный интервал второго, значит можно заподозрить кратность главных компонент.

Количество главных компонент можно посчитать с помощью Критерия Кайзера (отбираются факторы с собственными числами > 1, в данном примере их получается 3).

Так как собственные числа корреляционной матрицы являются дисперсиями главных компонент, то оценка уровня информативности первых трёх главных компонент составляет I3(z(x)) = , что превышает требуемый уровень 55%. I2(z(x)) =



Так же можно использовать Критерий каменистой осыпи, для этого в Statistica строится графическое представление собственных чисел («График осыпи» в русской Statistica).



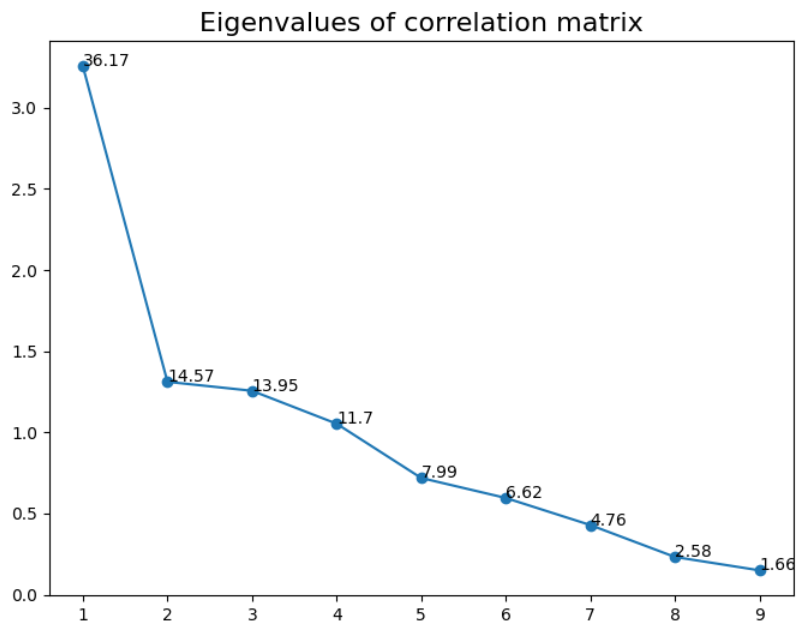
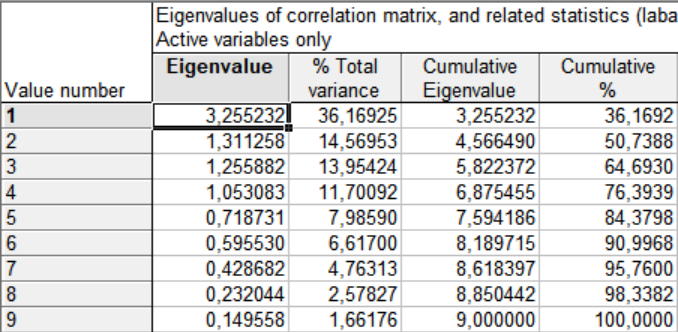
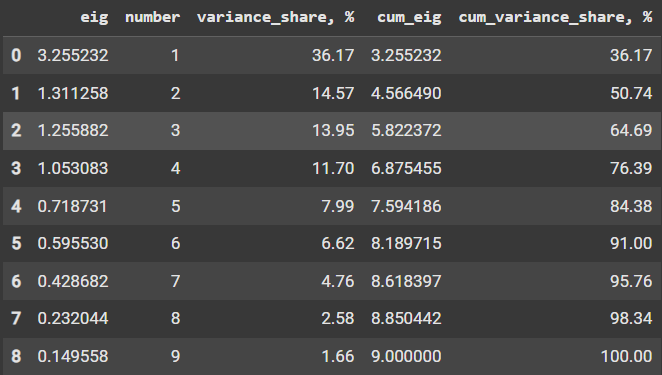


Рисунок 3 – График собственных чисел

Следует найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только "факториальная осыпь". В соответствии с этим критерием можно оставить в этом примере 2 главные компоненты.

Таблица 2 – Вклады главных компонент в суммарную дисперсию исходных признаков, рассчитанные в пакете Statistica





В первом столбце таблицы приведены оценки собственных чисел, в третьем столбце – накопленные значения собственных чисел, во втором и в четвертом столбцах – относительный вклад каждой главной компоненты в суммарную дисперсию и накопленный относительный вклад соответственно.

Таблица 3 – Результаты расчета собственных векторов корреляционной матрицы

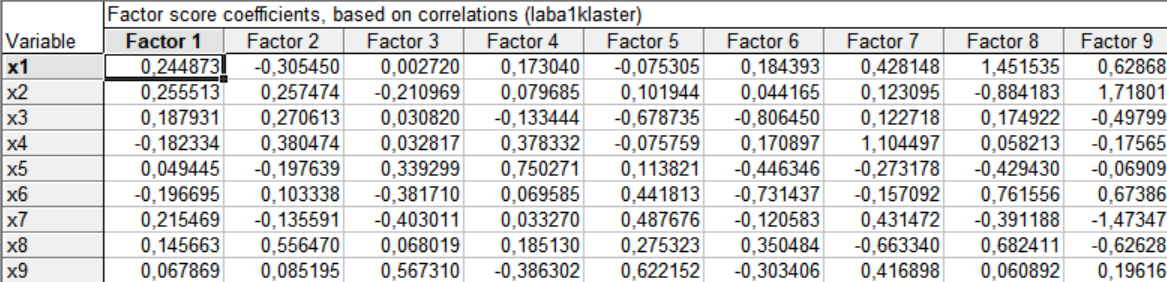
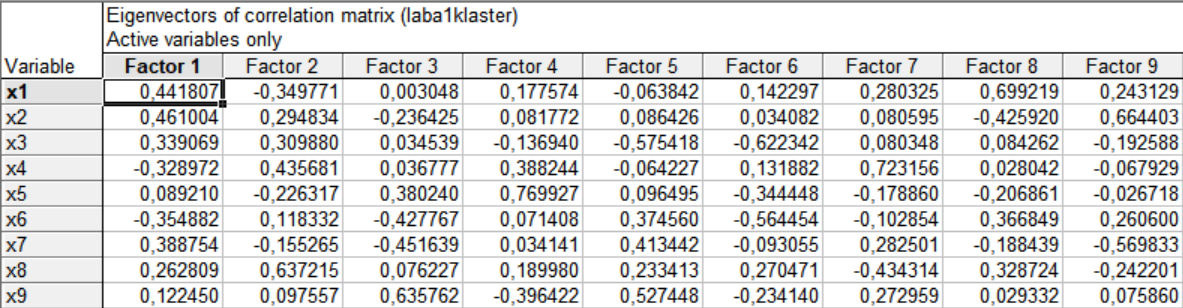
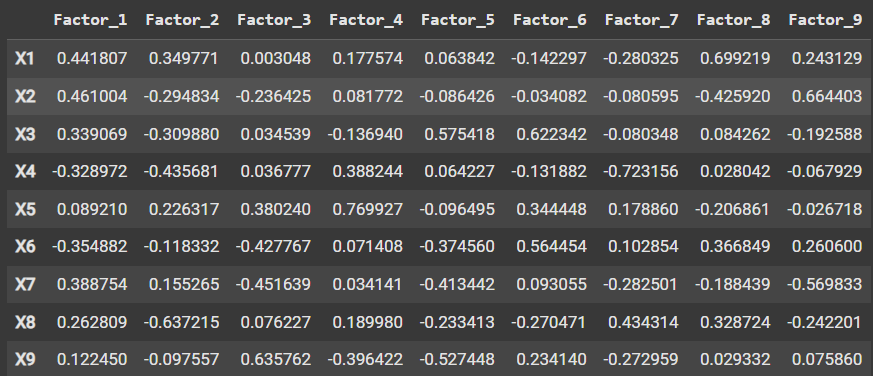


Таблица 4 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков





При снижении размерности признакового пространства до трех главных компонент следует рассматривать только три первых столбца матрицы U.

Главные компоненты связаны с центрировано-нормированными исходными признаками следующими линейными комбинациями:

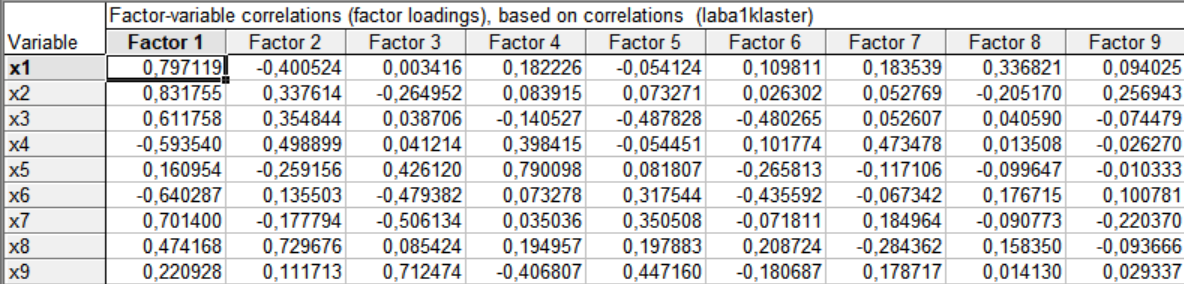
*z1* = 0,44x1\*+0,46x2\*+0,33x3\*-0,32x4\*+0,09x5\*-0,35х6\*+0,39x7\*+0,26x8\*+0,12х9\*

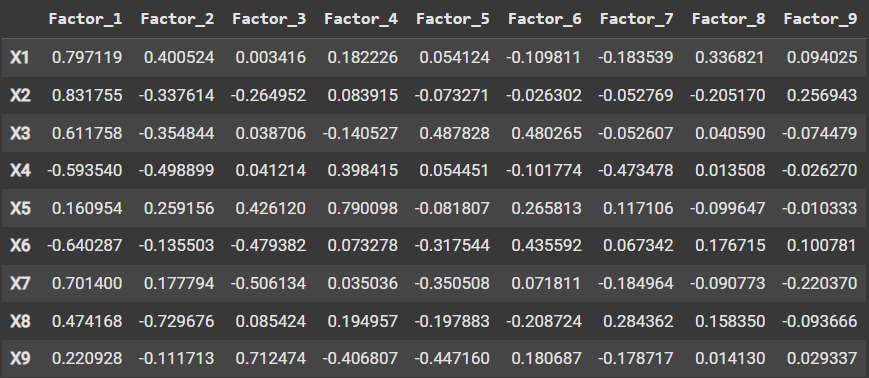
*z2* = -0,35x1\*+0,29x2\*+0,31x3\*+0,44x4\*-0,23x5\*+0,11х6\*-0,15x7\*-0,64x8\*-0,09х9\*

*z3* = 0,003x1\*-0,24x2\*+0,03x3\*+0,04x4\*+0,38x5\*-0,43х6\*-0,45x7\*+0,08x8\*+0,64х9\*

# **Матрица нагрузок**

Таблица 5 – Результаты расчета элементов 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 пятью главными компонентами не наблюдается тесной связи (не имеется значений >0,55). Это подтверждает правильность выделения только трех первых главных компонент. Матрица нагрузок имеет вид:

Первая главная компонента тесно связана (коэффициент корреляции больше 0,55) с 4 исходными признаками *x1* – брачность, *х2 -* средний размер назначенных пенсий, *x3 -* заболеваемость на 1000 человек населения, *x7 -* отребительские расходы в среднем на душу населения тыс. Рублей. И тесно отрицательно с *х6 -* число населения на одну больничную койку. Поэтому первую главную компоненту можно интерпретировать как «уровень здравоохранения и брачности».

Вторая главная компонента тесно положительно связана с признаками *x8* – количество врачей на 1000 человек населения. Поэтому вторую главную компоненту можно интерпретировать как «обеспеченность врачами».

К третьей главной компоненте можно отнести признаки *х9* – число спортивных сооружений на 1000 человек и охарактеризовать как «обеспеченность спортивными сооружениями». *х5* – число посещений музея на 1000 человек населения относим к четвертой компоненте, охарактеризуем как «посещаемость музея». Центрировано-нормированные исходные признаки связаны с центрировано- нормированными главными компонентами f1, f2, f3 следующими выражениями:

*x1\* =* 0,79*f1  f2+*0,003*f3;*

*x2\* =*0,83 *f1 +* 34*f2*26*f3;*

*x3\* =* 0,61*f1 f2+04f3;*

*x4\* = f1 + f2+f3;*

*x5\* =* 0,16*f1  f2+*0,43*f3;*

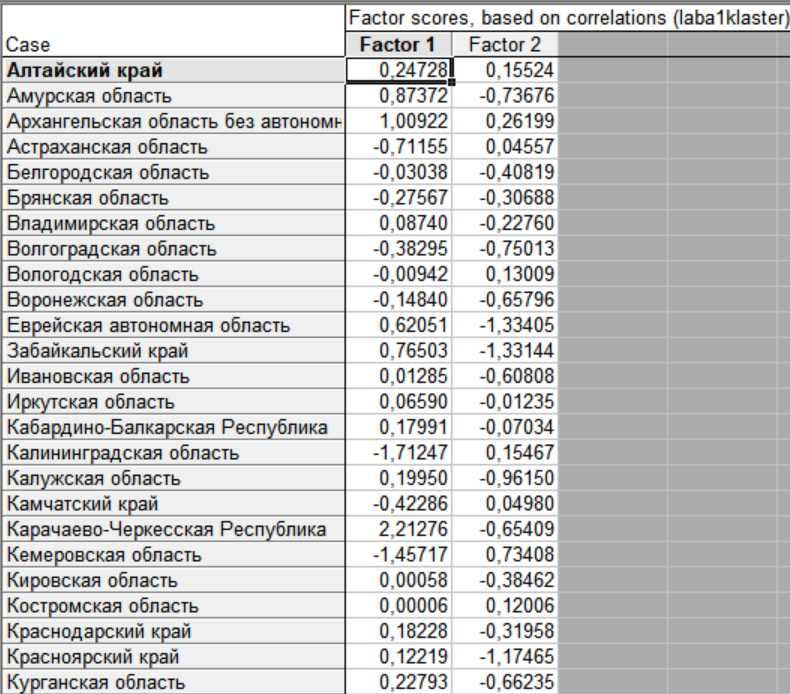
*x6\* = f1 + f2-*0,48*f3;*

*x7\* =* 0,70*f1 f2*0,51*f3;*

*x8\* =* 0,47*f1* 73*f2+*0,08*f3;*

*x9\* =* 0,22*f1* 11*f2+*0,71*f3;*

Рисунок 4 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент





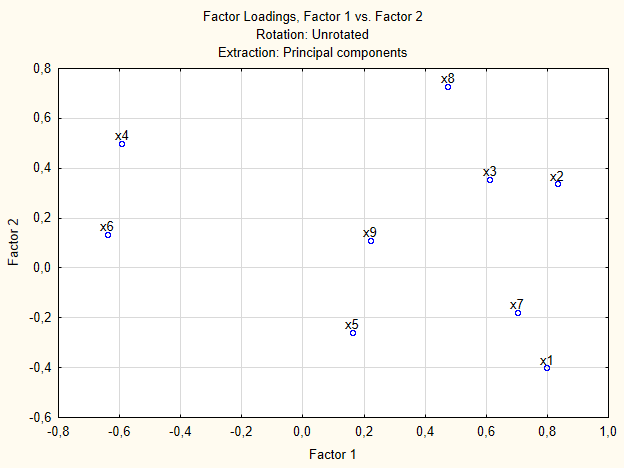


Рисунок 7 – Распределение признаков в пространстве первых двух главных компонент

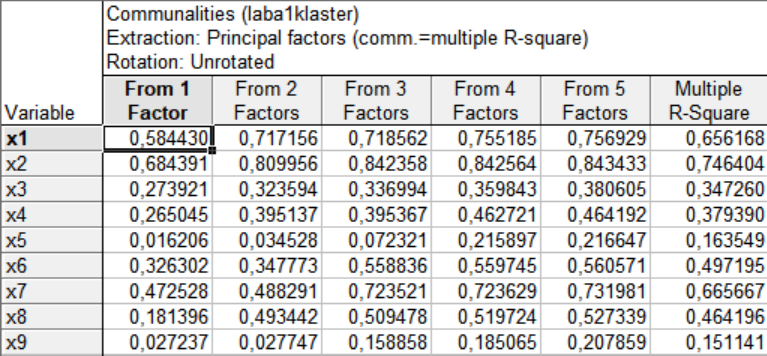
# **Метод главных факторов**

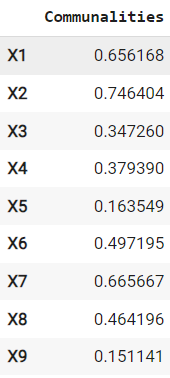
Оценками общностей в данном алгоритме будут служить квадраты оценок множественных коэффициентов корреляции , т.е.:

, .

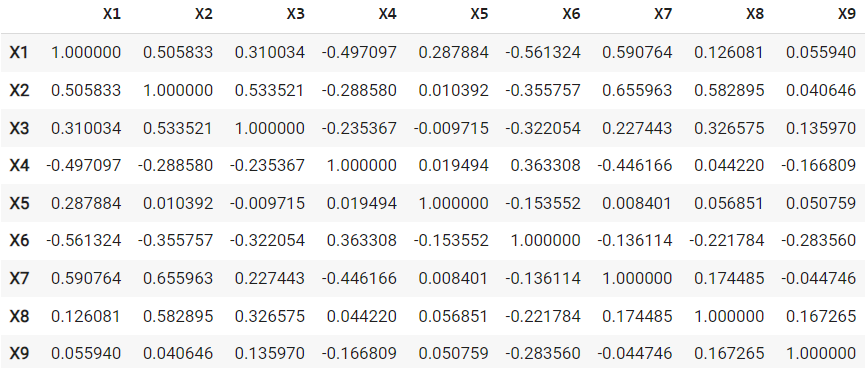
Оценки общности представлены ниже:

Таблица 6 – Результаты расчета общностей для двух главных факторов





В первом, втором столбцах таблицы содержатся вклады одного и двух факторов в дисперсию признаков. Оценки общностей приведены в третьем столбце таблицы. На основе матрицы парных коэффициентов корреляции и оценок общностей можно составить оценки редуцированной матрицы



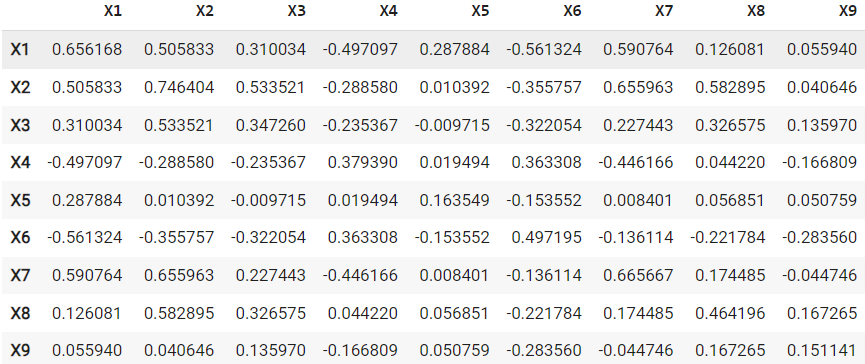
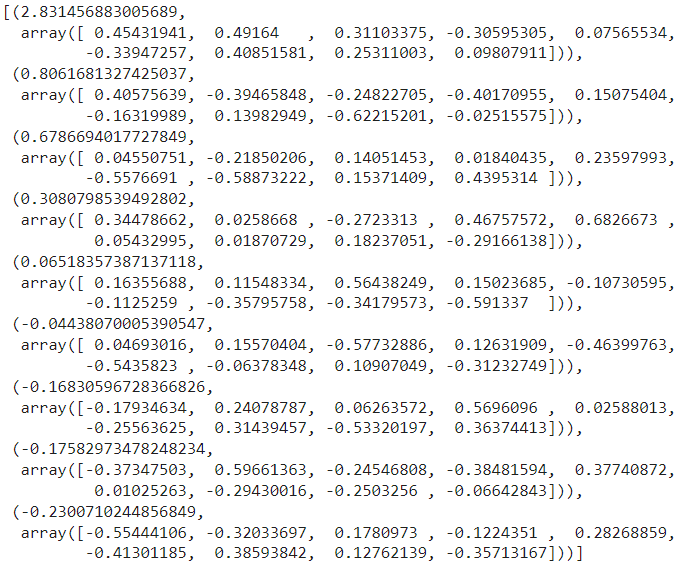
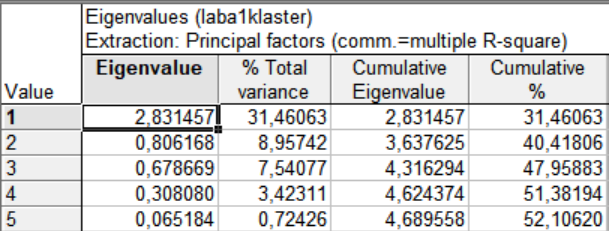
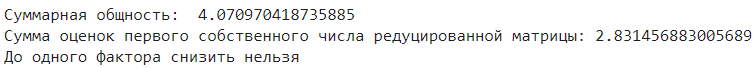


Таблица 7 – Вклад двух главных факторов в суммарную дисперсию исходных признаков









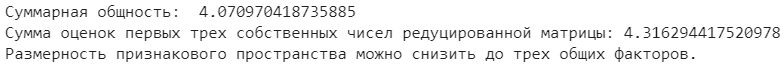


Таблица 8 – Результаты расчета общностей для трех главных факторов

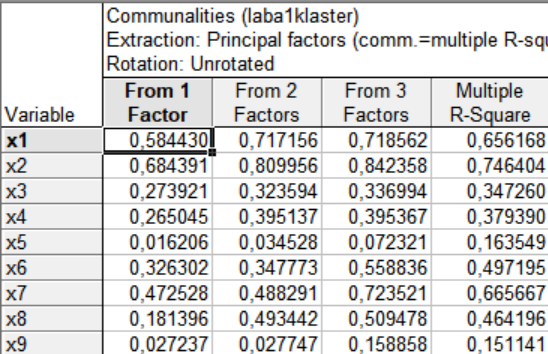
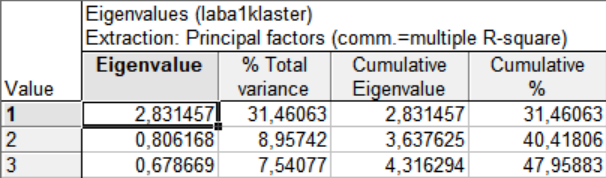
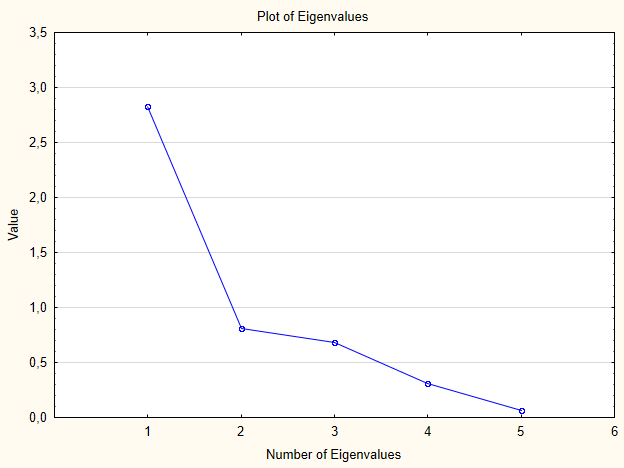


Таблица 9 – Вклад трех главных факторов в суммарную дисперсию исходных признаков



Сумма оценок трех собственных чисел редуцированной матрицы будет в этом случае больше суммарной общности (

Вклад трех главных факторов в суммарную дисперсию исходных признаков (в дисперсию процесса) составляет 47,96%.



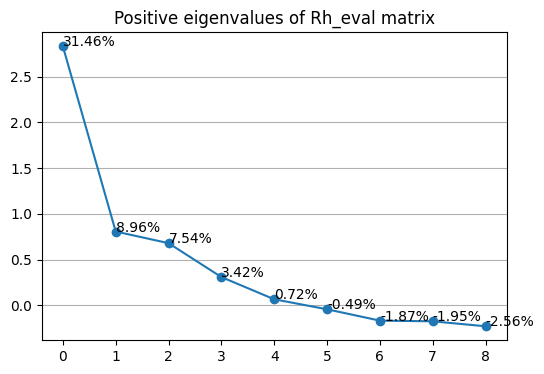


Рисунок 8 – График собственных значений

Таблица 10 – Весовые коэффициенты при общих факторах

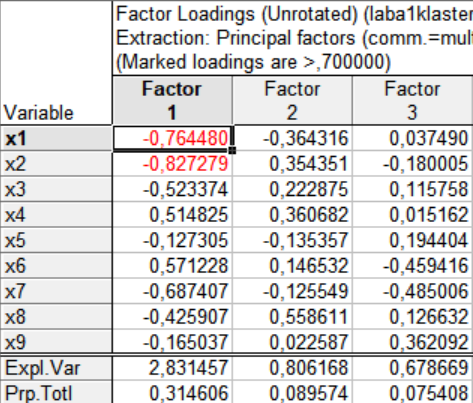
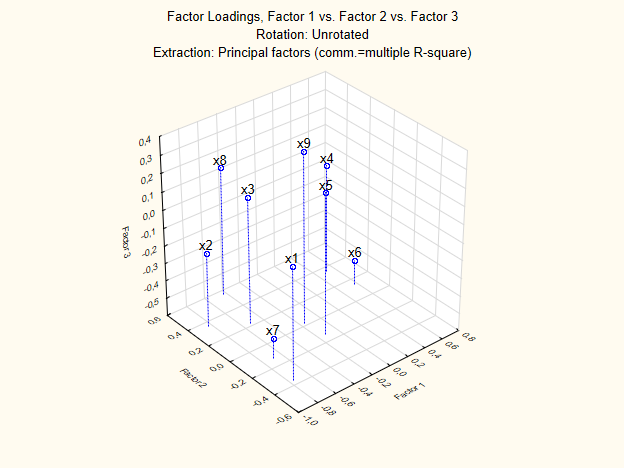
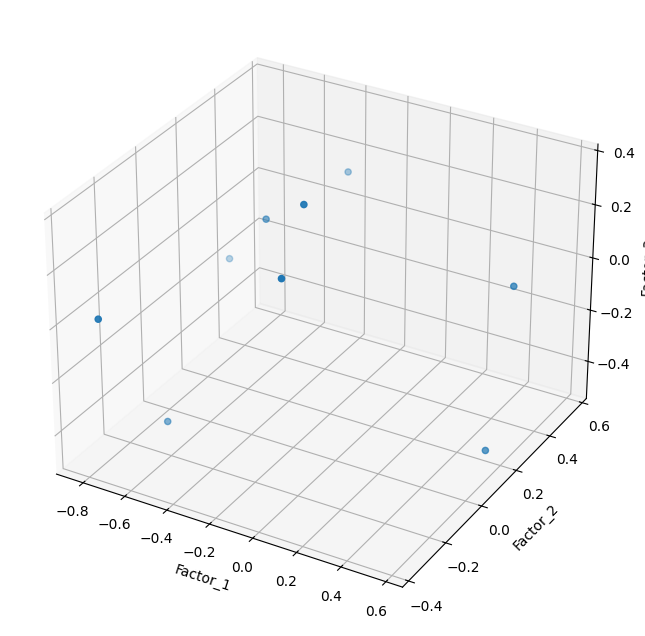




Рисунок 9 – Расположение исходных признаков на плоскости, образованной главными факторами





Попробуем упростить структуру главных факторов с помощью вращения. Для оценки структуры обобщённых факторов выберем критерий «Квартимакс», т.к. он показал наилучший результат среди других критериев.

Таблица 11 – Весовые коэффициенты факторов после вращения

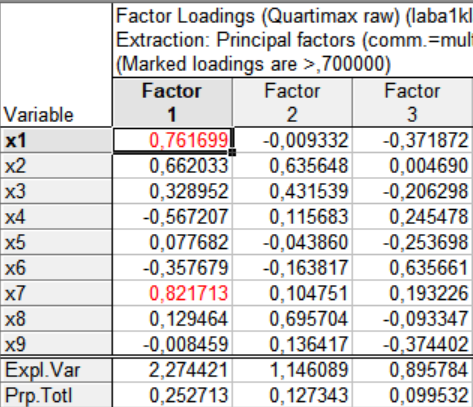
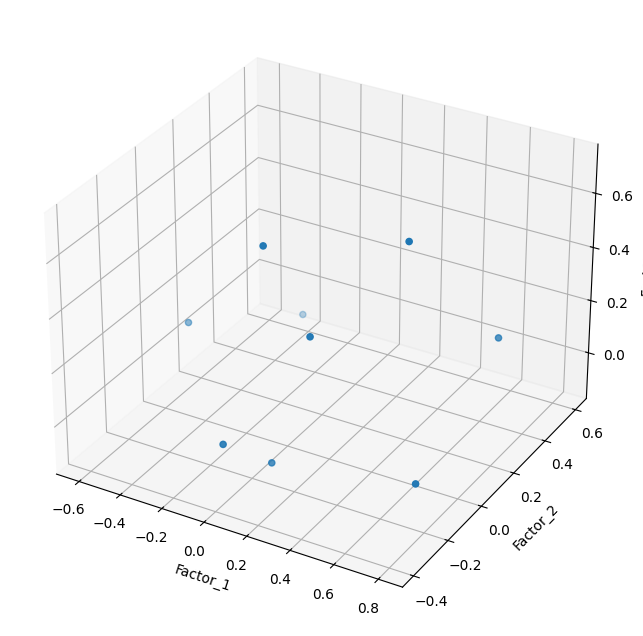
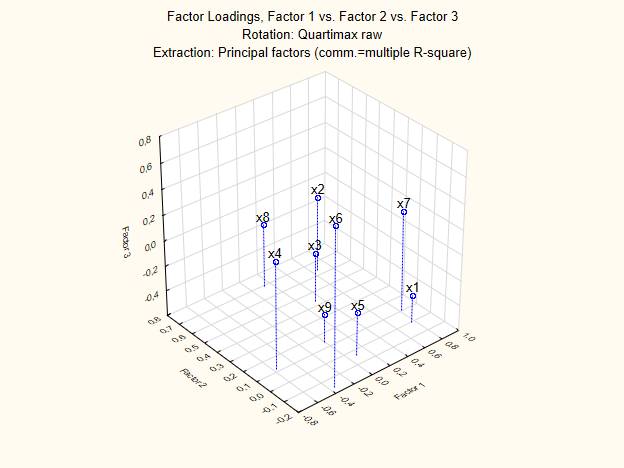




Рисунок 10 – Расположение исходных признаков на плоскости, образованной главными факторами после вращения





Матрица факторных нагрузок после вращения имеет вид:

|  |  |  |
| --- | --- | --- |
| 0,761699 | -0,009332 | -0,371872 |
| 0,662033 | 0,635648 | 0,004690 |
| 0,328952 | 0,431539 | -0,206298 |
| -0,567207 | 0,115683 | 0,245478 |
| 0,077682 | -0,043860 | -0,253698 |
| -0,357679 | -0,163817 | 0,635661 |
| 0,821713 | 0,104751 | 0,193226 |
| 0,129464 | 0,695704 | -0,093347 |
| -0,008459 | 0,136417 | -0,374402 |

|  |  |  |
| --- | --- | --- |
| 0,584430 | 0,717156 | 0,718562 |
| 0,684391 | 0,809956 | 0,842358 |
| 0,273921 | 0,323594 | 0,336994 |
| 0,265045 | 0,395137 | 0,395367 |
| 0,016206 | 0,034528 | 0,072321 |
| 0,326302 | 0,347773 | 0,558836 |
| 0,472528 | 0,488291 | 0,723521 |
| 0,181396 | 0,493442 | 0,509478 |
| 0,027237 | 0,027747 | 0,158858 |

X1 – брачность (число браков на 1000 человек)

X2 - средний размер назначенных пенсий

X3 – заболеваемость на 1000 человек населения

X4 - процент безработных мужчин

X5 - число посещений музея на 1000 человек населения

X6 - число населения на одну больничную койку

X7 – потребительские расходы в среднем на душу населения тыс. рублей

X8 – количество врачей на 1000 человек населения

X9 – число спортивных сооружений на 1000 человек

Вывод:

Сравнивая матрицы А и В можно сделать вывод, что структура обобщённых факторов после вращения значительно улучшилась. Первый фактор тесно связан (коэффициент корреляции > 0,7) с одним исходным признаком: *х9 -* Среднедушевые денежные доходы населения. Следовательно, первый главных фактор можно характеризировать как «Среднедушевые доходы».

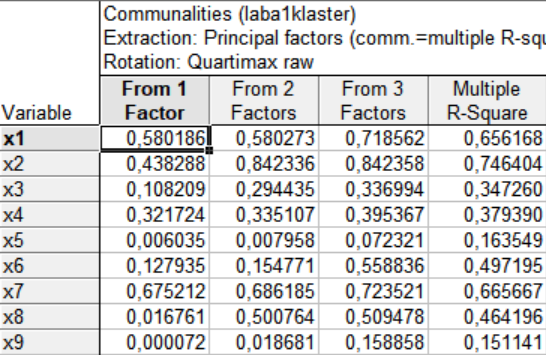
Второй фактор тесно не связан ни с какими исходными признаками:

Третий фактор также тесно не связан с остальными признаками*.*

Далее представлена таблица вкладов факторов в дисперсию исходных признаков. Элементы первого столбца таблицы равны квадратам соответствующих элементов первого столбца матрицы В, элементы второго столбца – сумме квадратов соответствующих элементов первого и второго столбцов матрицы В, элементы третьего столбца – сумме квадратов элементов первого, второго и третьего столбцов матрицы В. Таким образом, в третьем столбце представлены оценки общностей, рассчитанные по матрице В, а в четвертом столбце – оценки общностей, рассчитанные по формуле:

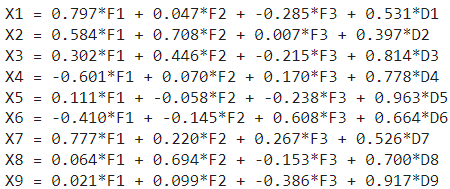
, 

Таблица 12 – Вклады трех факторов в дисперсию признаков



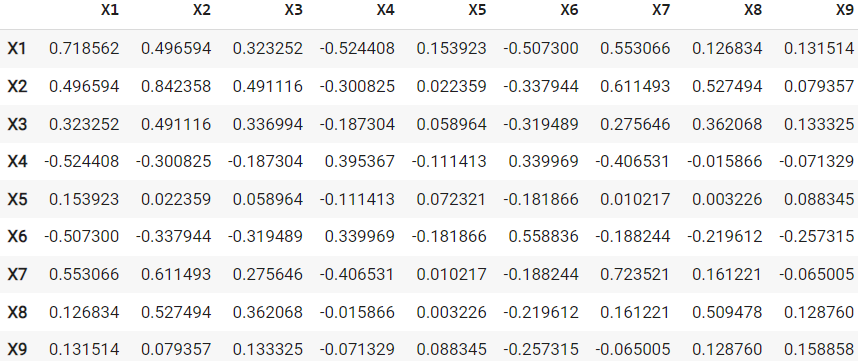
Т.к. исходные признаки процентированы и пронормированы, а главные факторы не коррелированы между собой, то оценки характерностей можно рассчитать следующим образом:

Центрировано-нормированные исходные признаки связаны с главными и характерными факторами следующими выражениями: ():



Оценка редуцированной матрицы парных коэффициентов корреляции, рассчитанная по матрице нагрузок В, и оценка остаточной матрицы парных коэффициентов корреляции представлены ниже:

Таблица 13 – Оценка редуцированной матрицы



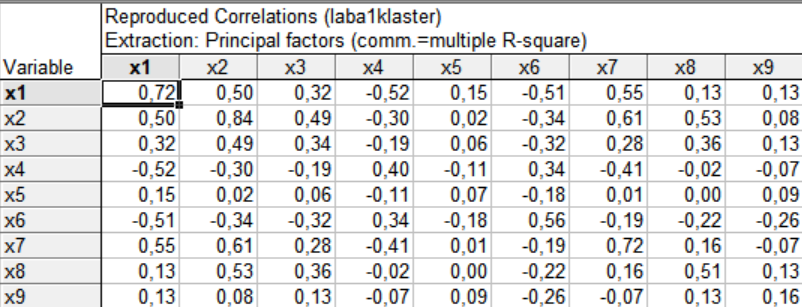
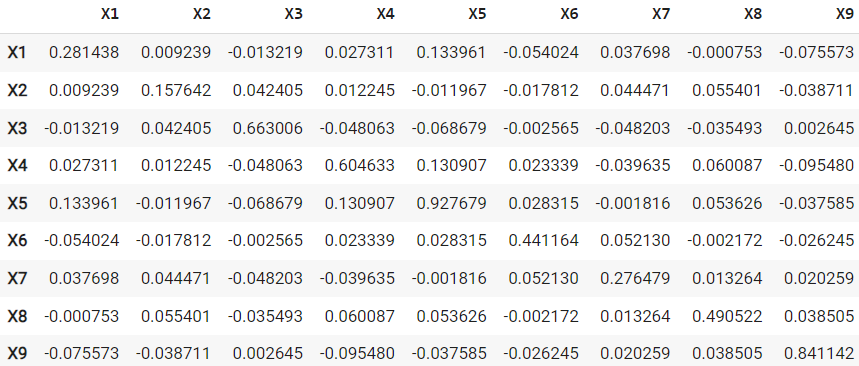
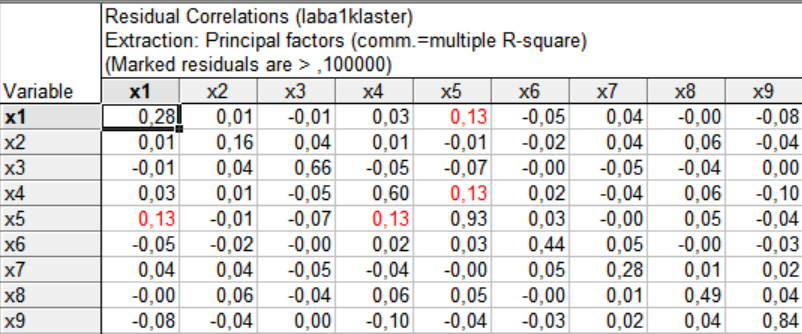
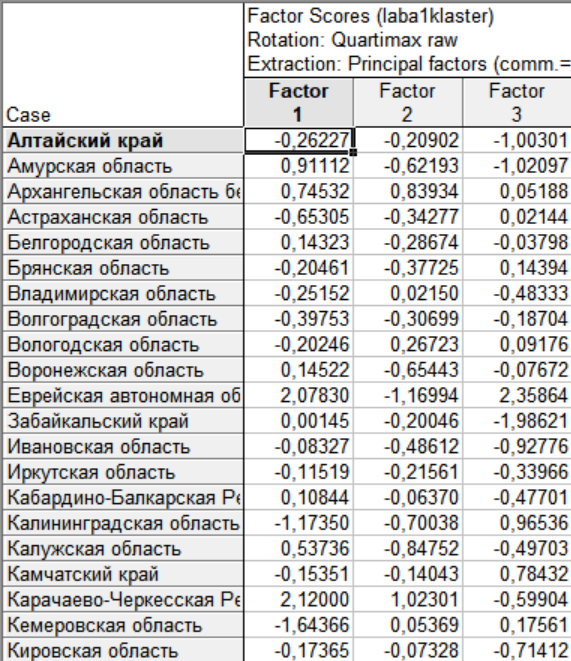


Таблица 14 – Оценка остаточной матрицы парных коэффициентов корреляции





На главной диагонали матрицы, представленной в табл.12, расположены оценки характерностей .



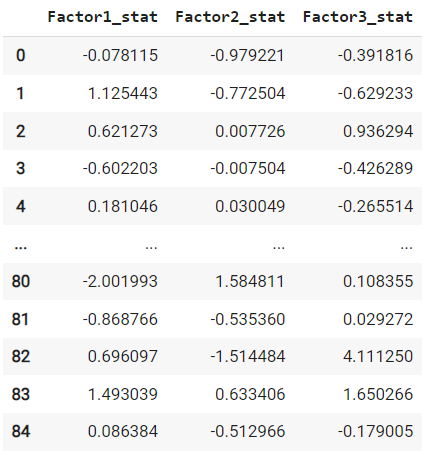


Рисунок 11 – Индивидуальные значения общих факторов

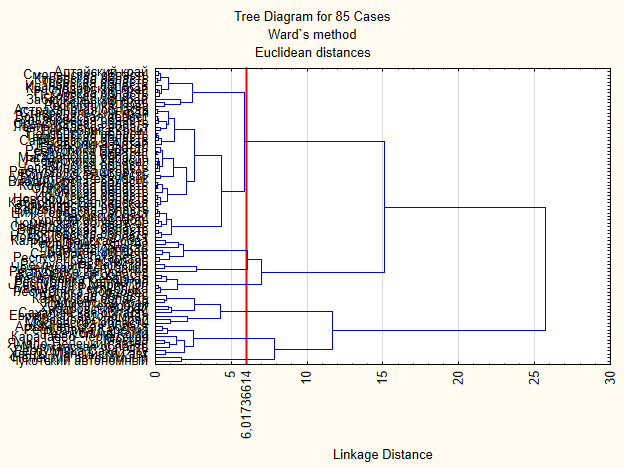


Рисунок 5 – Дендрограмма по Уорду по оставшимся двум главным компонентам

Таблица 6 – Результаты классификации субъектов РФ методом Уорда

|  |  |  |
| --- | --- | --- |
| **Номер**  **кластера** | **Кол-во**  **объектов** | **Состав класса** |
| кластер 1  {S11} | 6 | Кемеровская область Республика Калмыкия Республика Марий Эл Республика Мордовия Республика Северная Осетия – Алания Чувашская Республика |
| кластер 2  {S12} | 2 | Ненецкий автономный округ Чукотский автономный округ |
| кластер 3  {S13} | 7 | Калининградская область Камчатский край Липецкая область Республика Адыгея Республика Татарстан Самарская область Севастополь |
| кластер 4 {S14} | 9 | Архангельская область без автономного округа Карачаево-Черкесская Республика Москва Мурманская область Республика Карелия Республика Коми Республика Саха (Якутия) Ханты-Мансийский автономный округ – Югра Ямало-Ненецкий автономный округ |
| кластер 5 {S15} | 3 | Республика Дагестан Республика Ингушетия Чеченская Республика |
| кластер 6 {S16} | 9 | Амурская область Еврейская автономная область Калужская область Красноярский край Московская область Приморский край Санкт-Петербург Сахалинская область Хабаровский край |
| кластер 7 {S17} | 49 | Алтайский край Астраханская область Белгородская область Брянская область Владимирская область Волгоградская область Вологодская область Воронежская область Забайкальский край Ивановская область Иркутская область Кабардино-Балкарская Республика Кировская область Костромская область Краснодарский край Курганская область Курская область Ленинградская область Магаданская область Нижегородская область Новгородская область Новосибирская область Омская область Оренбургская область Орловская область Пензенская область Пермский край Псковская область Республика Алтай Республика Башкортостан Республика Бурятия Республика Крым Республика Тыва Республика Хакасия Ростовская область Рязанская область Саратовская область Свердловская область Смоленская область Ставропольский край Тамбовская область Тверская область Томская область Тульская область Тюменская область без автономных округов Удмуртская Республика Ульяновская область Челябинская область Ярославская область |

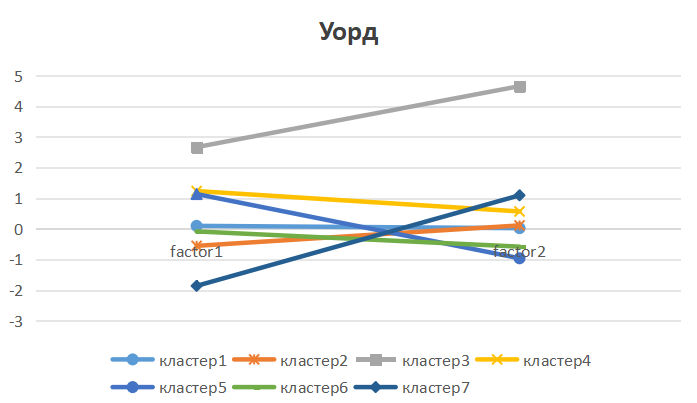


Диаграмма 1 – График средних значений кластеров по методу Уорда

Рисунок 6 - График рассеяния объектов по методу Уорда для 2 ГК(New)

Рисунок 6 - График рассеяния объектов по методу Уорда для 2 ГК(Old)

Таблица 7 – Результаты классификации субъектов РФ методом k-средних (2)

|  |  |  |
| --- | --- | --- |
| **Номер**  **кластера** | **Кол-во**  **объектов** | **Состав класса** |
| кластер 1  {S11} | 14 | Пермский край Алтайский край Вологодская область Иркутская область Кабардино-Балкарская Республика Костромская область Нижегородская область Орловская область Курская область Новгородская область Республика Башкортостан Тверская область Республика Коми Республика Саха (Якутия) |
| кластер 2  {S12} | 2 | Ненецкий автономный округ Чукотский автономный округ |
| кластер 3  {S13} | 16 | Липецкая область Удмуртская Республика Астраханская область Камчатский край Республика Бурятия Ставропольский край Ульяновская область Республика Адыгея Пензенская область Республика Марий Эл Республика Мордовия Тамбовская область Чувашская Республика Республика Алтай Калининградская область Республика Тыва |
| кластер 4 {S14} | 28 | Томская область Владимирская область Краснодарский край Оренбургская область Московская область Республика Татарстан Севастополь Брянская область Ростовская область Рязанская область Волгоградская область Курганская область Ленинградская область Новосибирская область Псковская область Республика Крым Республика Хакасия Самарская область Саратовская область Тульская область Челябинская область Белгородская область Воронежская область Магаданская область Смоленская область Кировская область Ивановская область Омская область |
| кластер 5 {S15} | 11 | Амурская область Еврейская автономная область Приморский край Санкт-Петербург Хабаровский край Забайкальский край Красноярский край Калужская область Свердловская область Тюменская область без автономных округов Ярославская область |
| кластер 6 {S16} | 8 | Архангельская область без автономного округа Республика Карелия Москва Мурманская область Ханты-Мансийский автономный округ – Югра Ямало-Ненецкий автономный округ Карачаево-Черкесская Республика Сахалинская область |
| кластер 7 {S17} | 6 | Кемеровская область Республика Дагестан Республика Ингушетия Республика Калмыкия Республика Северная Осетия – Алания Чеченская Республика |

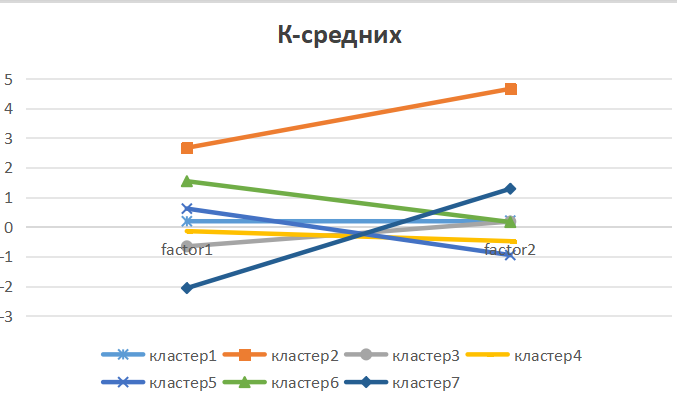
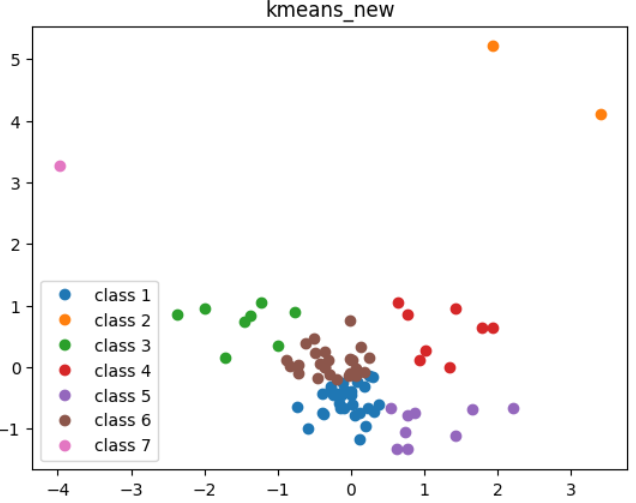


Диаграмма 2 - средние значения метода к-средних для 2 ГК



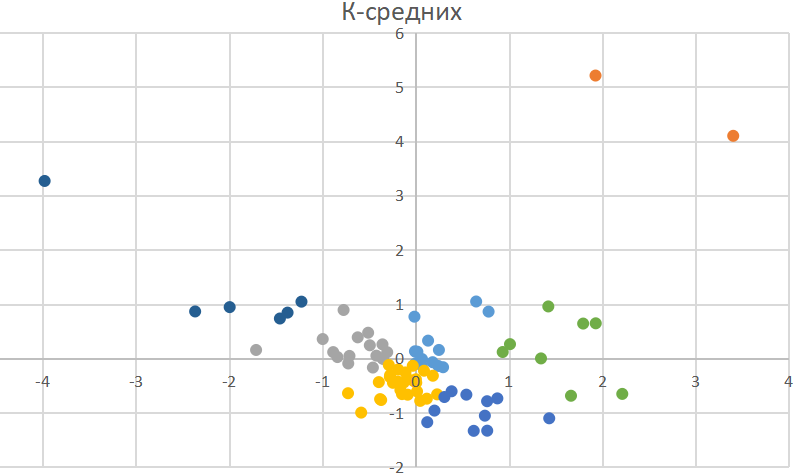
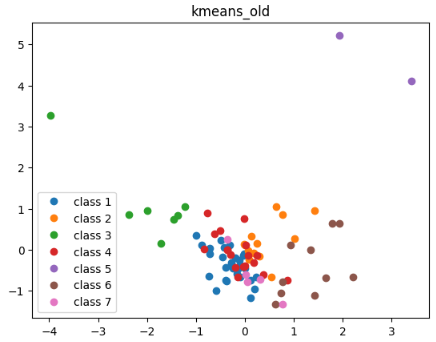


Рисунок 7 - График рассеяния объектов по методу К-средних для 2 ГК(New)



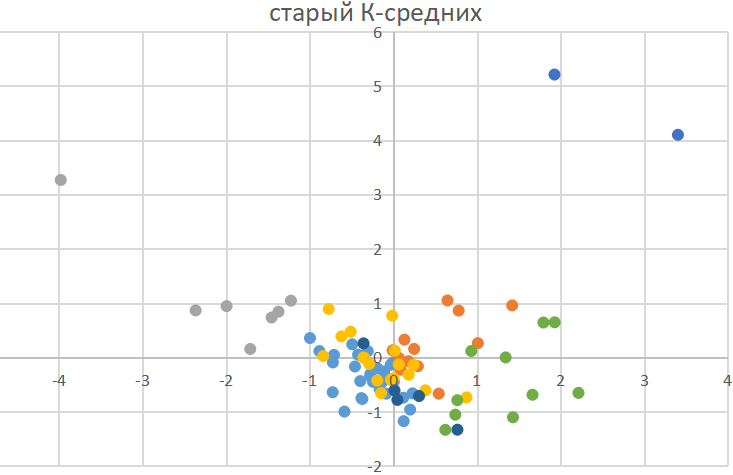


Рисунок 8 - График рассеяния объектов по методу К-средних для 2 ГК(Old)

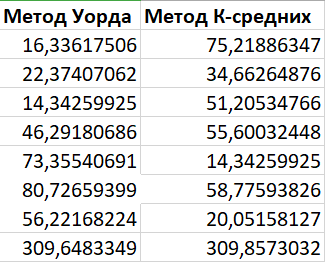
 

Рисунок 9 – Сводная таблица функционал качеств результатов классификации