Высшая школа экономики
(Научно-исследовательский университет)
Факультет компьютерных наук

Домашнее задание 2
по дисциплине: “Методы прикладной статистики”

Статистическое исследование доходов, расходов и сбережений населения

Студент: Рябыкин Алексей Сергеевич
Преподаватель: Сиротин Вячеслав Павлович
Оценка: _

Москва 2022

Оглавление

1. Постановка задачи

Выбранные данные: Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE).

Этапы работы:

1. Произвести отбор признаков для последующей классификации объектов с целью разделения общей совокупности на однородные подсовокупности. Произвести предварительный анализ этих признаков. Построить ортогональный базис из отобранных признаков методом главных компонент. Проанализировать возможность снижения размерности признакового пространства при сохранении в достаточной мере его информативности.
2. Произвести параметрическую классификацию объектов в отсутствие обучения методом расщепления смесей вероятностных распределений по предполагаемому наиболее информативному признаку либо по первой главной компоненте;
3. Разбить объекты на кластеры в пространстве главных компонент. Сравнить результаты кластеризации и классификации на основе декомпозиции смеси распределений;
4. Создать обучающую выборку из наиболее типичных представителей выделенных однородных групп объектов. Произвести классификацию объектов методом дискриминантного анализа. Произвести проверку правильности классификации методом кросс-валидации;
5. Построить в выделенных одним из способов однородных группах регрессионные модели по иным, чем используемые в классификации, признакам. Сравнить полученные модели и сделать выводы по проделанной работе.

Целью исследования можно назвать формирование понимания и формализацию закономерностей в финансовой и имущественной обеспеченности населения России.

Регионом исследования был выбран Москва.

2. Описание данных и выбранных признаков

Повтор из прошлой работы

Доходы, исходя из данных и международных принципов их дифференциации, могут быть представлены следующими видами:

Доходы	Виды доходов в исследовании
Первичные доходы	Оплата труда
	Продажа сельскохозяйственной продукции
	Личное хозяйство
	Предпринимательская деятельность
Собственность	Аренда недвижимости
	Проценты
Трансферты	Пенсии
	Стипендии
	Пособия
	Алименты
	Пособия по безработице
	Другие доходы
Прочие поступления	Продажа личного имущества
	Продажа недвижимости

Таблица 1 – Доходы и их виды

С другой стороны, среди расходов и имуществ населения можно выделить следующие группы для анализа:

Расходы	Примеры
Потребительские расходы	Продукты питания
Платежи и взносы	Налоги, связь, интернет
Приобретение имущества	Квартиры, гараж и прочее
Денежные сбережения	Вклады, сбережения в валюте

Таблица 2 – Расходы и их виды

Активы	Примеры
Реальные активы	Квартира, дача, машина
Оборотные активы	Одежда, наличные
Финансовые активы	Арендное имущество

Таблица 3 – Типы имущества

Для решения поставленных задач были выбраны немного отличные от предыдущего домашнего задания данные.

3. Предварительный анализ данных

Выбранные данные содержат результаты опроса 413 респондентов из Москвы по практически 1.5 тысячам вопросам различного характера.

Столбец	Описание	Количество п. з.	Доля п. з.
ze13.31b	Траты на медикаменты	85	20.6%
ze13.32b	Траты на моющие средства	113	27.4%
ze13.33b	Траты на хоз товары	86	20.8%
ze9.9b	Траты на интернет	63	15.3%
ze9.8b	Траты на мобильную связь	13	3.1%

Продолжение на следующей странице

Столбец	Описание	Количество п. з.	Доля п. з.
zc1.1	Стоимость жильзя	10	2.4%
zb1.o	Численность домохозяйств	30	7.3%
zc6	Полезная площадь жилья	0	0.0%
zc5	Жилая площадь	0	0.0%
zf12_a	Как долго смогли бы жить только за счет сбережений?	0	0.0%

Эта таблица может отличать от соответствующей в прошлом задании, потому что я обозначил все значения из множества [“ЗАТРУДНЯЮСЬ ОТВЕТИТЬ”, “ЖИЛЬЕ ПРОДАТЬ НЕВОЗМОЖНО”, “ЖИЛЬЕ НЕ ПОДЛЕЖИТ ПРОДАЖЕ”, “ОТКАЗ ОТ ОТВЕТА”, “НЕТ ОТВЕТА”] за пропуски, чего не сделал в прошлый раз. Считаю так более репрентативно. Для столбцов траты на интернет, пропуски были заполнены нулями, для остальных средним и модой для количественных и качественного признаков, соответсвенно.

Замечание

Описание некоторых столбцов здесь и впредь было изменено для адекватного отображения таблиц, рисунков, диаграмм.

Проведем предварительный анализ выбранных количественных признаков, рассчитав для каждого выборочные характеристики центральной тенденции и разброса:

	mean	std	min	25%	50%	75%	max
Траты на медикаменты	2429.48	3134.18	0.0	410.00	1500.00	3000.00	30000.0
Траты на моющие средства	647.68	528.08	30.0	380.00	647.47	647.47	5000.0
Траты на хоз товары	749.06	872.78	40.0	350.00	647.47	800.00	10000.0
Траты на интернет	615.45	276.38	0.0	490.00	550.00	614.89	2000.0
Траты на мобильную связь	1199.62	1273.79	0.0	500.00	1000.00	1500.00	20000.0
Стоимость жилья	9478285.7	3941801.8	2000000	6750000	9486781.61	12000000	20000000
Численность домохозяйств	2.55	1.44	1.0	1.00	2.00	3.00	9.0
Полезная площадь	53.28	15.32	16.5	42.60	52.30	60.20	113.0
Жилая площадь	34.11	12.16	11.0	26.55	32.00	42.00	77.0
Денежный доход	88784.06	65912.56	4000.0	45000.00	70172.00	113500.00	631000.0
Траты на питание	23377.69	12675.36	3500.0	15000.00	20000.00	30000.00	100000.0

Ниже приведены распределения и диаграммы рассеивания для выбранных для факторного анализа признаков:

Столбец	Описание
ze13.31b	Траты на медикаменты
ze13.32b	Траты на моющие средства
ze13.33b	Траты на средства личной гигиены
ze9.9b	Траты на Интернет
ze9.8b	Траты на мобильную связь
ze11	Траты на квартиру (аренда, коммунальные услуги)
ze4	Траты на питание (дома и вне дома)

Таблица 5 – Выбранные для факторного анализа столбцы

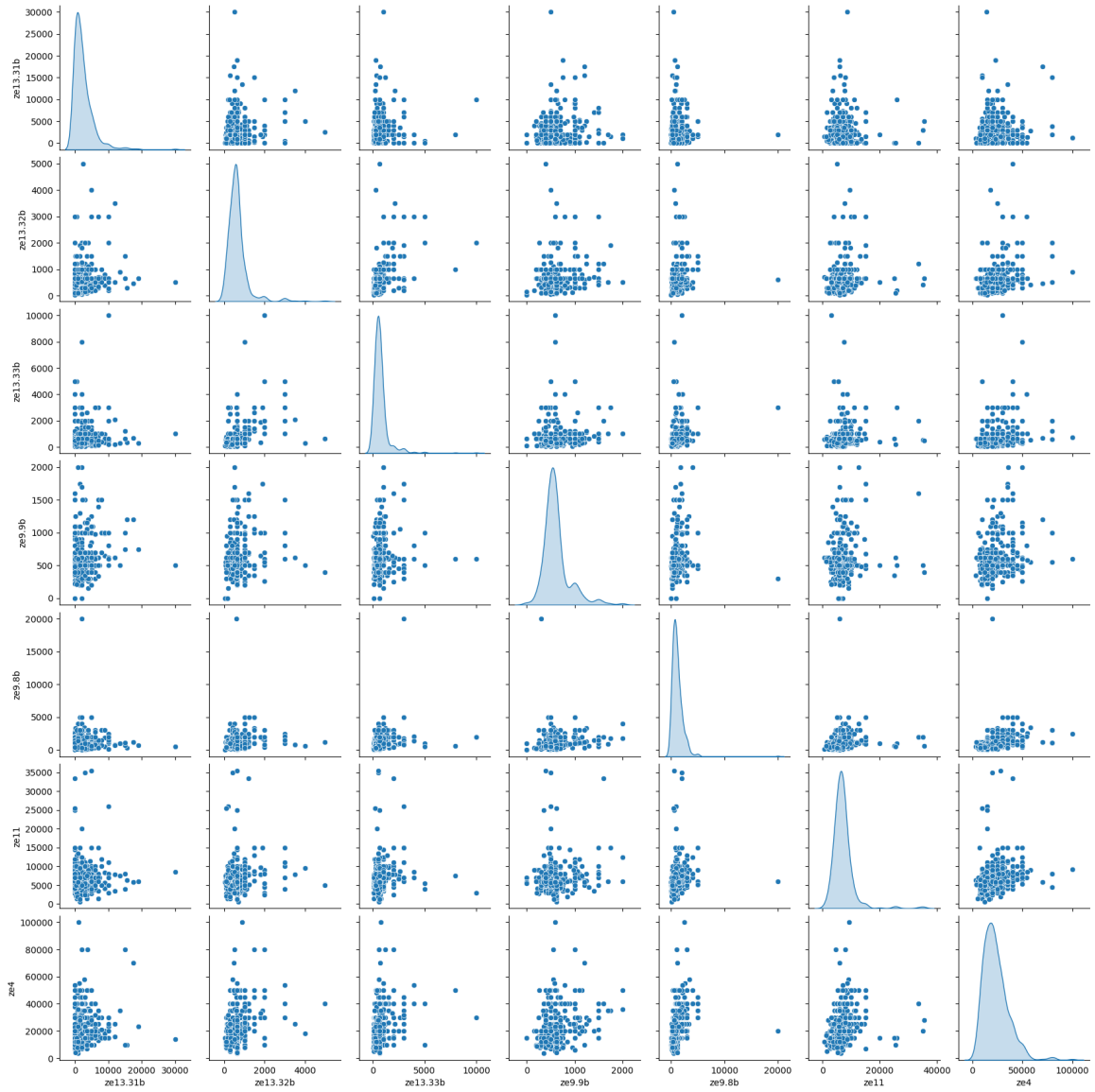


Рис. 1 – Непараметрические распределения и диаграммы рассеивания для каждого из выбранных для факторного анализа признака

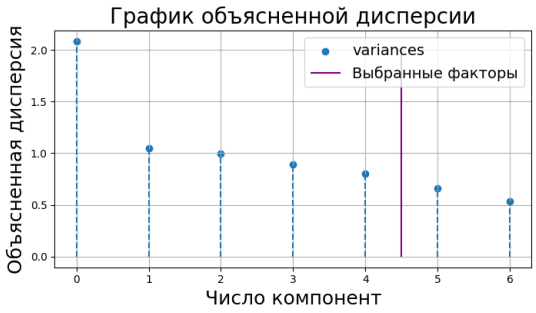
4. Метод главных компонент. Ортогональный базис. Снижение размерности.

Поиск главных компонент с помощью SVD разложения. Для получения n-главных компонент необходимо взять первые n сингулярных векторов, соответствующих n сингулярным значениям в отсортированном по убыванию порядке.

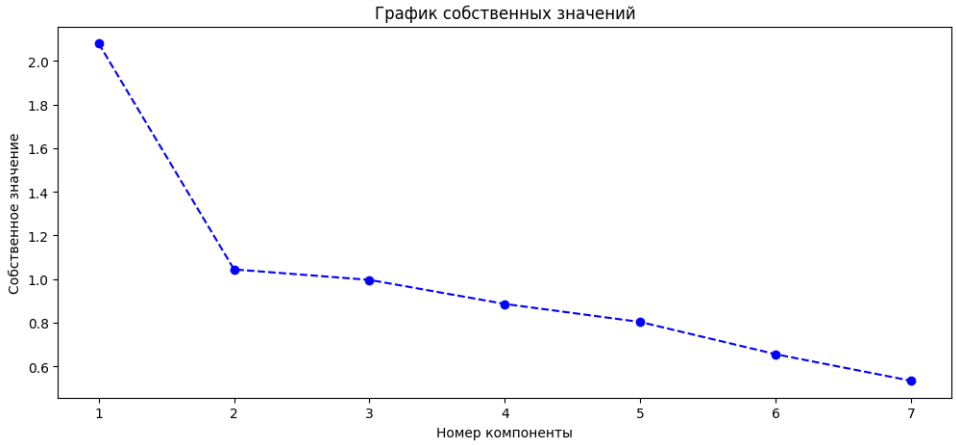
$$A = U\Sigma V^T$$

где U и V – ортогональные матрцицы с ортонормированными собственными векторами матриц AA^T и A^TA, соответственно. Σ – матрица сингулярных значений, каждое из которых равно квадратному корню собственного значения любой из матриц AA^T или A^TA (собственные значения их равны).

В силу того, что изначальные данные обладали высокой дисперсией, они были прошкалированы (Standard Scaler). После чего был применен метод главных компонент. В результате были выбраны 4 компоненты, сохраняющие немногим более 78% дисперсии.



Ниже представлен график собственных значений:



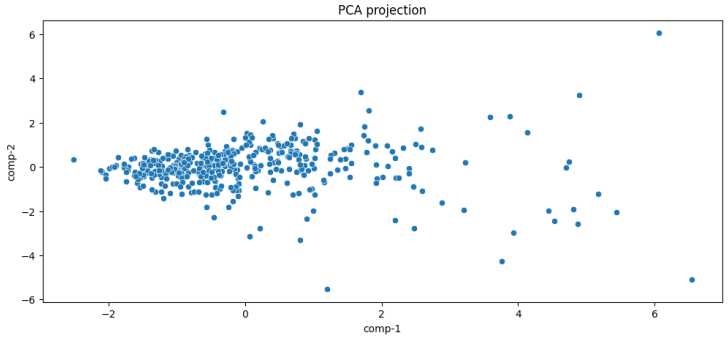
Посмотрим на матрицу факторных нагрузок:

Номер компоненты / Признак	Factor 1	Factor 2	Factor 3	Factor 4
Траты на медикаменты	0.141197	-0.008138	0.135708	0.00865
Траты на моющие средства	0.575594	0.106344	0.043137	0.197001
Траты на средства личной гигиены	0.687187	0.269792	0.043559	-0.0059069
Траты на Интернет	0.029865	0.243704	0.640133	-0.0019427
Траты на мобильную связь	0.113735	0.493452	0.199075	0.018566
Траты на квартиру (аренда, коммунальные услуги)	0.077945	0.223865	0.220932	0.054067
Траты на питание (дома и вне дома)	0.150570	0.627606	0.106392	0.589123

Попробуем интерпретировать компоненты:

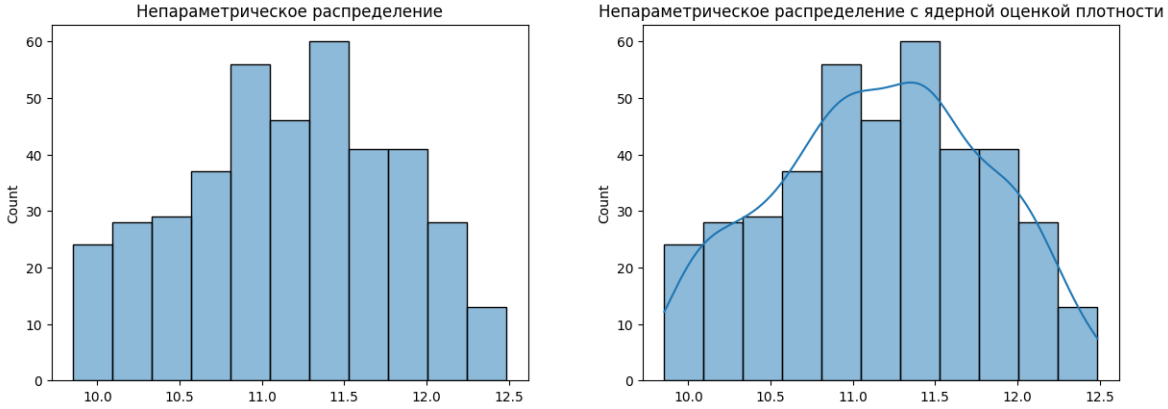
- Первая компонента может показывать траты на предметы первой необходимости для домохозяйств;
- Вторая компонента в некой степени описывает ежемесячные траты;
- Третья компонента показывает траты на квартиру;
- Четвертую компоненту интерпретировать сложнее, можно попробовать как траты на питание и всё, что с этим связано (мытье посуды, например).

В силу того, что новое пространство сохраняет немалый процент дисперсии – можно говорить о том, что снижение размерности позволительно. Посмотрим на данные после снижения размерности (в рамках двух первых компонент):



5. Расщепление смеси вероятностных распределений

Для параметрической классификации был выбран признак логарифма денежного дохода. Построим непараметрическое эмпирическое распределение и непараметрическое распределение с ядерной оценкой плотности вероятности:



Здесь возможно композиция сразу нескольких распределений, попробуем композицию 3 колокообразных распределений. Воспользуемся EM-алгоритмом для нахождения весов и параметров распределения, чтобы представить искомое в виде их линейной комбинации вида:

$$f(x) = \sum_{j=1}^{k} q_j f(x, \theta_j).$$

Опишем модель. Одномерная модель (как в нашем случае) представляет собой следующее расщепление плотности:

$$p(x) = \sum_{i=1}^{K} q_i f(x|\mu_i, \sigma_i).$$

Причем

$$\sum_{i=1}^{K} q_i = 1.$$

Для многомерного случая:

$$p(\vec{x}) = \sum_{i=1}^{K} q_i f(\vec{x}|\mu_i, \Sigma_i)$$

$$f(\vec{x}|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} \exp\left(-0.5(\vec{x} - \mu_i)^T \Sigma_i^{-1} (\vec{x} - \mu_i)\right)$$

Как уже было сказано, для такого обучения без прецедентов используется EM-алгоритм, состоящий из нескольких шагов:

- Инициализация:
  - Случайно выбрать примеры без замены из датасета $X = \{x_1, \dots, x_N\}$ и присвоить их значения оценкам средних $\{\hat{\mu}_i\}, \forall i = [1, \dots, K]$;
  - Для всех оценок дисперсии присвоить значение выборочной дисперсии:

$$\hat{\sigma}_1^2, \dots, \hat{\sigma}_K^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2,$$

    где среднее – выборочное среднее:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

  - Пропорции распределений оценить равномерным распределением:

$$\hat{q}_1, \dots, \hat{q}_2 = \frac{1}{K}.$$

- Expectation (E) шаг: $\forall i, k$ оценить вероятность, что i-ый сэмпл сгенерирован из k-ой компоненты:

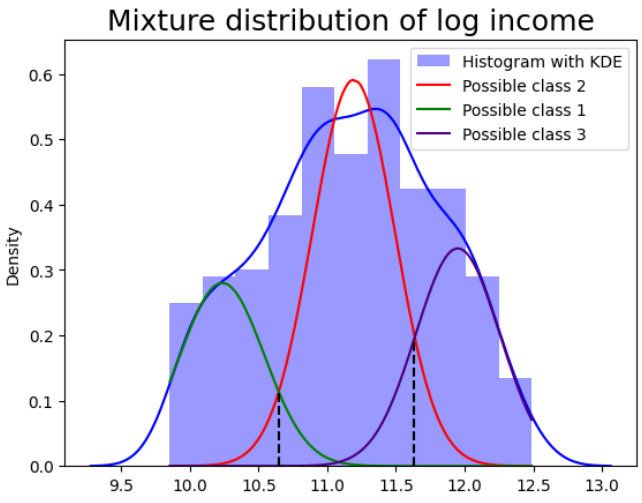
$$\hat{\gamma}_{ik} = p(C_k|x_i, \hat{q}, \hat{\mu}, \hat{\sigma}) = \frac{\hat{q}_k f(x_i|\hat{\mu}_k, \hat{\sigma}_k)}{\sum_{j=1}^{K} \hat{q}_j f(x_i|\hat{\mu}_j, \hat{\sigma}_j)}.$$

- Maximization (M) шаг: Пересчитать оценки в соответствии с посчитанными вероятностями принадлежности сэмплов к классам (компонентам, кластерам):

$$\forall k: \quad \hat{q}_k = \sum_{i=1}^{N}\frac{\hat{\gamma}_{ik}}{N}; \qquad \hat{\mu}_k = \frac{\sum_{i=1}^{N}\hat{\gamma}_{ik}x_i}{\sum_{i=1}^{N}\hat{\gamma}_{ik}}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_{ik}(x_i - \hat{\mu}_k)^2}{\sum_{i=1}^{N}\hat{\gamma}_{ik}}.$$

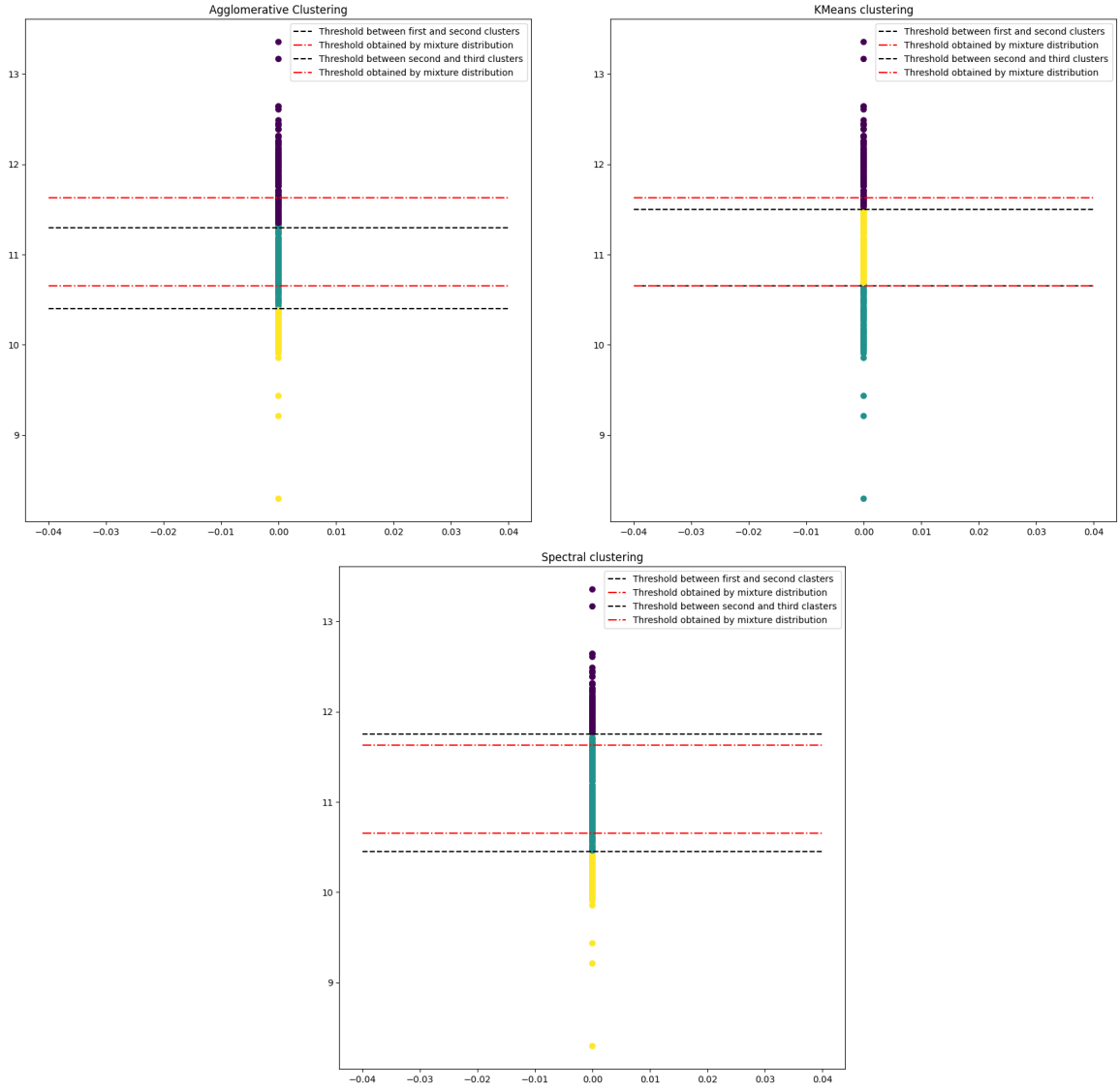
В итоге, получаем:



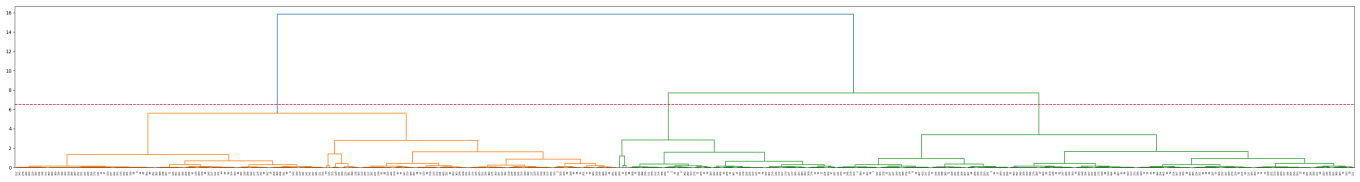
Имеем три класса и границы их отделимости. Далее была произведена разметка выборки в соотвествии с полученными классами, чтобы сравнивать результаты с методами кластеризации и последующими регрессионными моделями.

6. Кластеризация. KMeans, Hierarchiral, Spectral

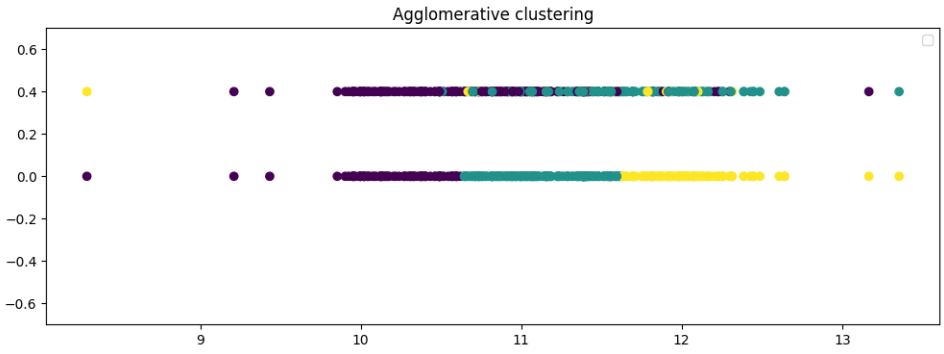
Для сравнения кластеризация производилась с помощью методом К средних, иерархической и спектральной кластеризаций. Для кластеризации по признаку логарифма дохода результаты схожи с расщеплением распределений и приведены ниже:

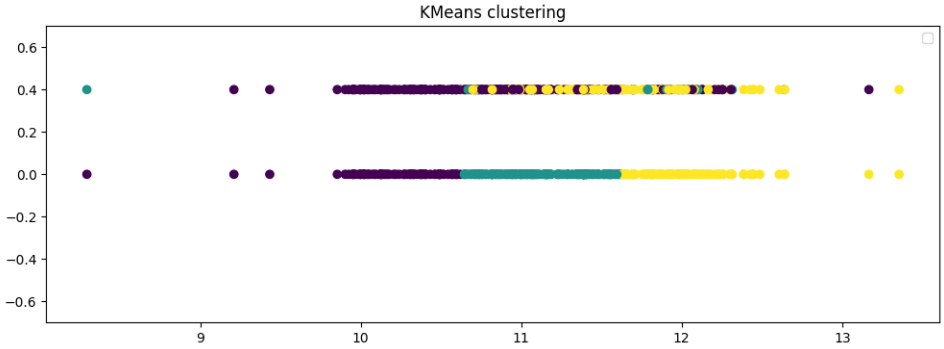


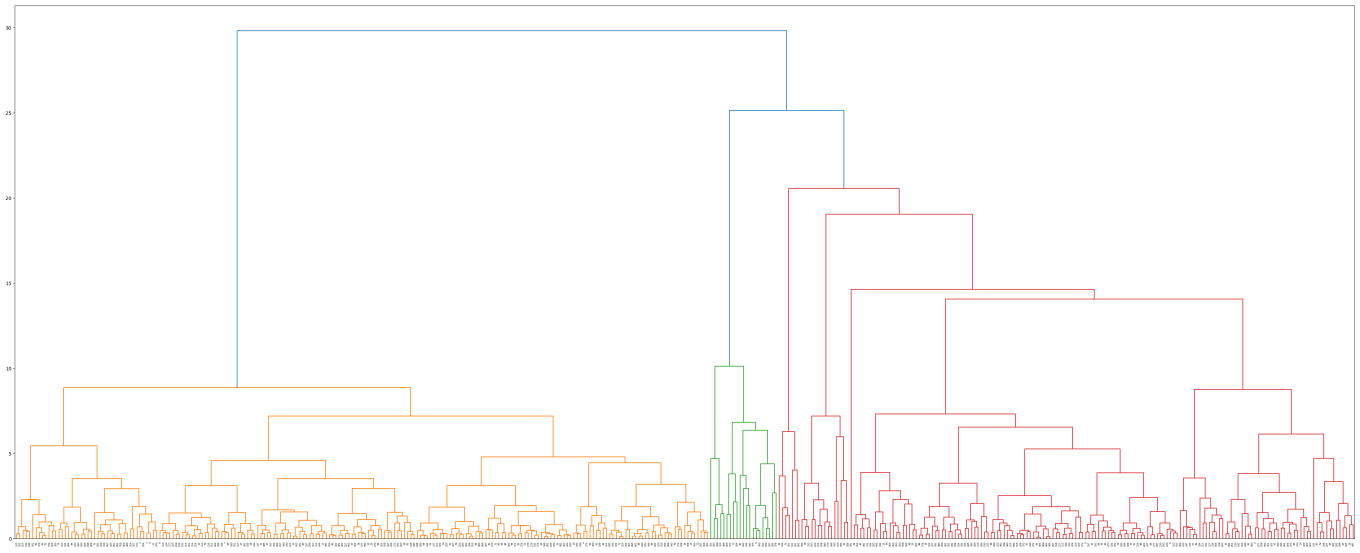
Дендрограмма для иерархической кластеризации представлена ниже:



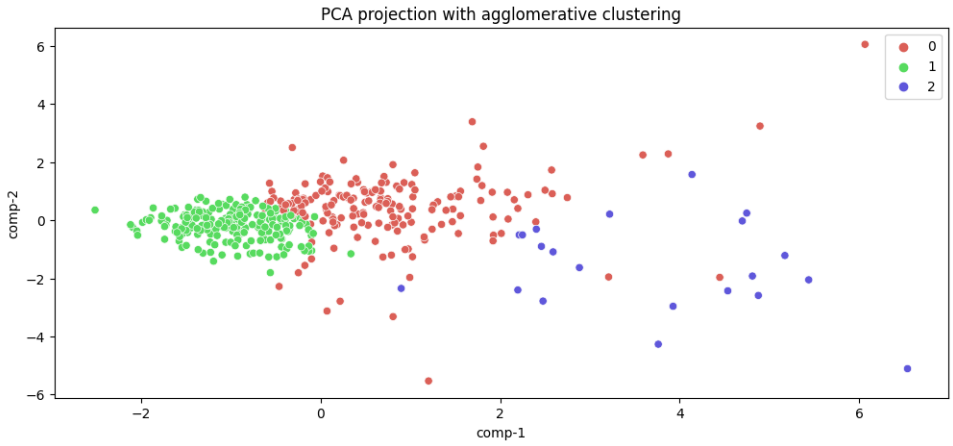
При кластеризации по объектам в пространстве главных компонентам присутствуют значительные отличия:

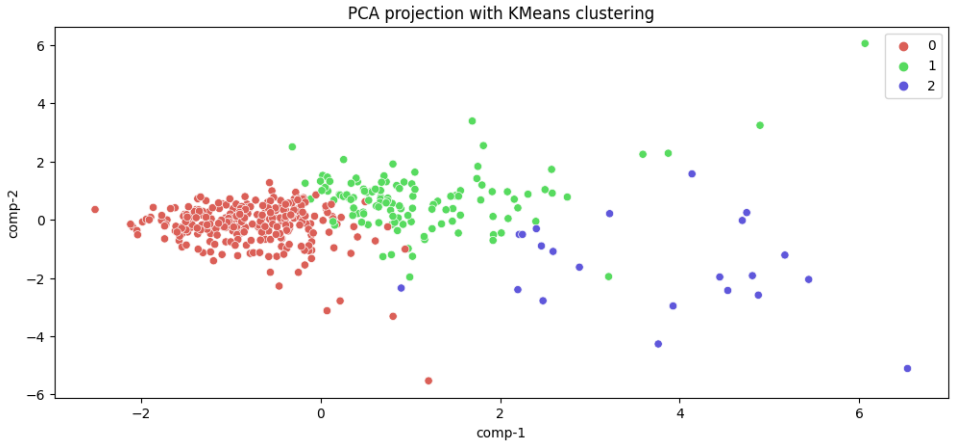




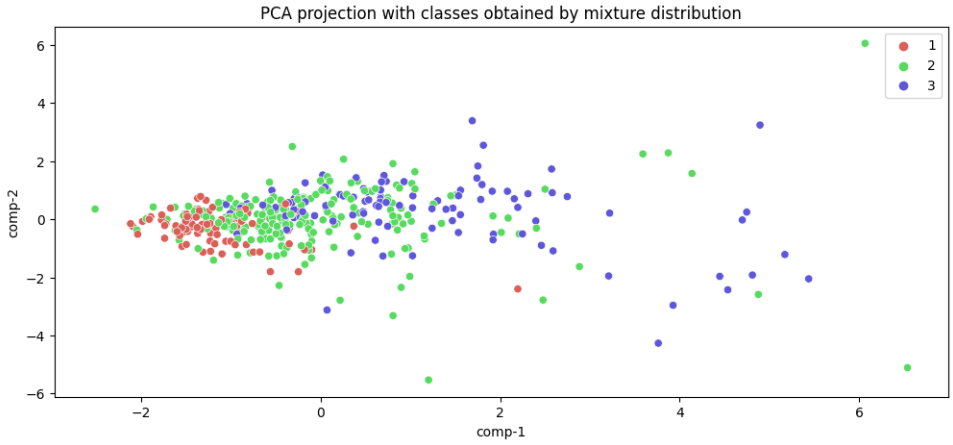


Однако визуализация в рамках двух главных компонент, показывает о наличии структуры и валидности таких разбиений на кластеры:





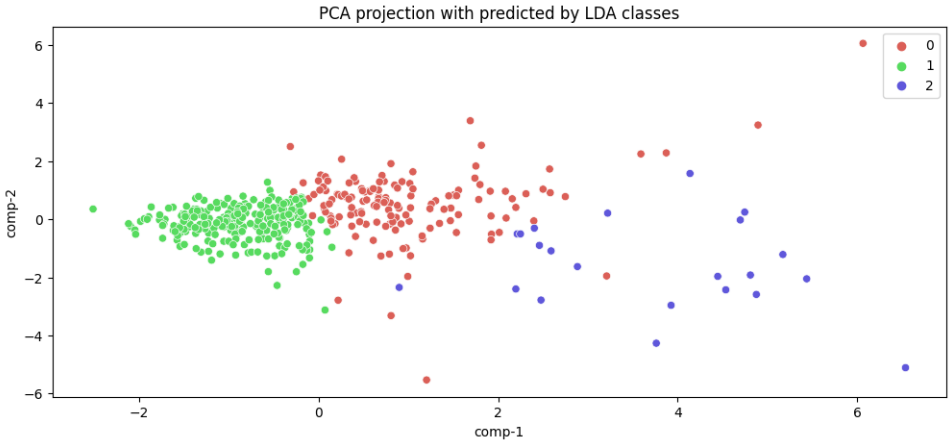
Конечно, в случае с расщеплением распределения денежных доходов всё не так хорошо, но он явно и не содержался в выбранных признаках для факторного анализа. Наличие, даже приведенной ниже структуры, уже говорит о ее существовании.

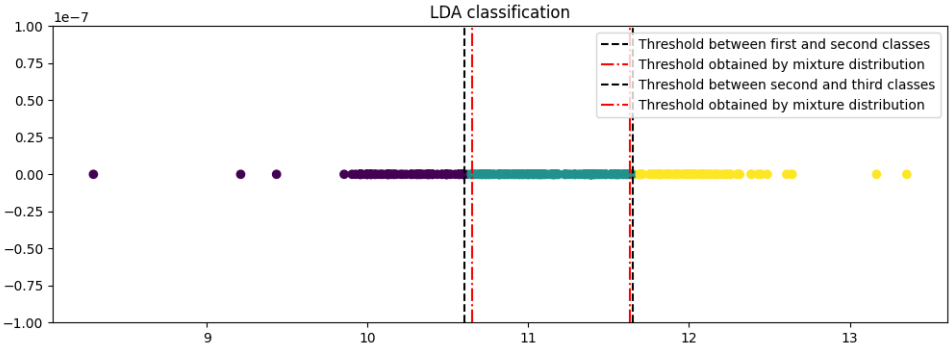


7. Классификация методом дискриминантного анализа

Так же разметим выборки в соответствии с остальными методами кластеризации. Построим кросс-валидацию двумя методами: LeaveOneOut и стандартный KFoldation. Посмотрим на метрики точность и F1 меру (weighted обобщенную на многоклассовую классификацию).

Метрика \ Таргет	income		pca_cols	
			Hierarchical	KMeans
	LOO	KFoldation	LOO	KFoldation
accuracy	0.98	0.9878	0.898	0.893
f1 (weighted)	0.98	0.9876	0.898	0.8884





8. Регрессионные модели

Выбранные признаки для регрессии:

Столбец	Обозначение	Описание
zc1.1	x_1	Какова сегодня приблизительно рыночная стоимость такого жилья, как Ваше?
zb1.o	x_2	Сколько человек, включая Вас, в Вашей семье, домохозяйстве?
zc6	x_3	Какова общая полезная площадь жилья у Вашей семьи, то есть сумма площадей жилых комнат, кухни, ванной, туалета, прихожей, кладовых и тому подобного в квартире (доме)?
zc5	x_4	Какую жилую площадь занимает Ваша семья? Сколько квадратных метров составляет площадь только жилых комнат?
zf12_a	x_5	Если все члены Вашей семьи лишатся всех источников дохода, как долго Ваша семья сможет материально жить так же, как сейчас, т.е. не уменьшая расходов, только за счет денежных сбережений, ничего не продавая из имущества?

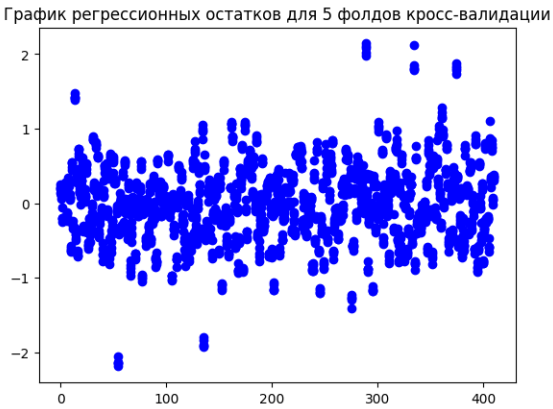
Последний признак является категориальным, обработка была сделана двумя способами (OneHotEncoding и LabelEncoder). Регрессия была построена с таргетом y на логарифм дохода, точность была посчитана с помощью метрик R2, MSE и MAE. Результаты приведены ниже. После этого в соответствии с границами из расщепления распределений, был произведен перевод в кластеры и посчитаны метрики f1 и точность. После всего этого, была запущена кросс-валидация, и проверены остатки регрессии на гомоскедастичность (устойчивость дисперсии) и на нормальность распределения на каждом фолде кросс-валидации. В результате отклоняем гипотезу о нормальности распределения регрессионных остатков, но не отвергаем гипотезу о гомоскедастичности. Модель устойчива, но таргет не способен был полностью объяснен выбранными признаками.

$$\hat{y} = 0.00000029x_1 + 0.3x_2 - 0.00032x_3 + 0.0031x_4 + 0.031x_5 + 10.2. \quad R2 = 0.344,\ MAE = 0.4018,\ MSE = 0.27.$$

После перевода в классы:

accuracy score: 0.6, f_1 score: 0.634

Регрессионные остатки:



Гомоскедастичность с помощью теста Левена не отвергается. p-value велик. По критериям Шапиро-Уилка и теста Харке-Бера распределение остатков регрессии не нормально (p-value сильно близок к нулю).