Национальный исследовательский университет «Высшая школа экономики»
Факультет компьютерных наук
Образовательная программа: Прикладная математика и информатика

Отчет по Модульному домашнему заданию №1
по майнору «Прикладной статистический анализ»

«Статистический анализ численности студентов по регионам Российской Федерации»

Работу выполнила
студентка 2 курса
Ульянова Александра Игоревна
Преподаватель:
Тихонова Арина Михайловна

Москва, 2023 г.

Содержание

Введение

Для проведения анализа в данной работе мной были выбраны следующие показатели:

1) Уровень безработицы по субъектам Российской Федерации на 2020 г., %
2) Доля внутренних затрат на исследования и разработки в валовом региональном продукте (ВРП) по субъектам Российской Федерации на 2020 г., %
3) Прирост высокопроизводительных рабочих мест по субъектам Российской Федерации за 2020 год, %
4) Индекс потребительских цен на все товары и услуги по субъектам Российской Федерации в 2020 году, в % к декабрю предыдущего года
5) Численность студентов, обучающихся по программам высшего образования очной формы обучения, по субъектам Российской Федерации на 2020 год, человек

Гипотеза исследования заключается в том, что существует сильная корреляция между 5 вышеперечисленными признаками. Вероятно, во время кластерного анализа регионы Российской Федерации поделятся на несколько групп в зависимости от их экономического положения. Предполагается, что регионы с низким уровнем безработицы также будут регионами с высокой долей затрат на исследования и разработки, в них будет большой прирост высокопроизводительных рабочих мест, индекс потребительских цен в них будет не очень большим, а студентов очной формы обучения будет много. Еще одна гипотеза состоит в том, что число студентов очной формы обучения зависит от четырех остальных факторов. Т.е. в регионах с плохим экономическим положением меньше людей предпочитают идти получать высшее образование, тем более очно, а вместо этого выбирают работать, возможно получая при этом заочное образование.

Цель исследования – проверить вышеуказанную гипотезу. Задачи исследования состоят в подборе данных, проведение предварительного, корреляционного и кластерного анализа данных, а также в подведении итогов и интерпретации полученных результатов.

Описание показателей:

Массив исходных данных представляет собой таблицу, в которой каждый столбец отвечает за один из пяти показателей, указанных в начале данного раздела, а каждая из 85 строк представляет каждый из субъектов Российской Федерации.

Приведем более подробное описание каждого из показателей:

1) Уровень безработицы - это показатель, который определяет долю экономически активного населения в возрасте 15 лет и старше, которое не имеет работы, но активно ищет ее в течение последнего месяца. Этот показатель может быть выражен как процент от общего числа экономически активного населения. Это непрерывный количественный показатель.

2) Доля внутренних затрат на исследования и разработки в валовом региональном продукте - это показатель, который измеряет объем финансирования исследований и разработок в определенном регионе, выраженный в процентах от общего объема валового регионального продукта (ВРП). Этот показатель используется для оценки уровня инновационности и технологического развития региона. Чем выше доля внутренних затрат на R&D в ВРП, тем больше регион инвестирует в научные исследования, технологические инновации и развитие инфраструктуры. Это непрерывный количественный показатель.
3) Прирост высокопроизводительных рабочих мест - это увеличение числа рабочих мест, которые связаны с производством высокотехнологичных продуктов и услуг, требующих высокой квалификации и инновационных знаний. Это непрерывный количественный показатель.
4) Индекс потребительских цен на все товары и услуги - это статистический показатель, который отражает изменение среднего уровня цен на все товары и услуги, которые приобретают средние домохозяйства в определенной стране или регионе. Это непрерывный количественный показатель.
5) Численность студентов, обучающихся по программам высшего образования очной формы обучения – дискретный количественный показатель.

В дальнейшем для краткости эти признаки будут обозначаться мной как признаки №1, №2, №3, №4 и №5 соответственно.

Данные были взяты с сайта Федеральной службы государственной статистики, а также с сайта Министерства образования и науки Российской Федерации. Выбор именно этих ресурсов был связан с тем, что они предоставляют официальные и самые новые данные, на которые опираются органы власти, научное сообщество и СМИ. Также они собирают информацию, которой зачастую нет в других государственных базах данных или в административных источниках.

Предварительный анализ данных

В ходе предварительного анализа данных вычислим для каждого из исследуемых нами признаков следующие характеристики: характеристики положения (среднее, моду, медиану), характеристики разброса (размах вариации, коэффициент вариации, дисперсию, стандартное отклонение), ранговые характеристики (квантили, децили), Z-преобразование, интерквантильный размах и некоторые прочие. Данные вычисления помогут нам лучше понять особенности данных, с которыми мы работаем, получить более четкое представление об исследуемых переменных.

Данные расчеты были проведены мной в программе Excel с использованием встроенных в нее математических функций. Более подробно формулы, которые были использованы, представлены ниже в Таблице 1.

Таблица 1. Формулы, использованные для расчета характеристик

Характеристика	Формула для вычисления с помощью встроенных функций Excel
Среднее арифметическое	=СРЗНАЧ([x_i])
Медиана	=МЕДИАНА([x_i])
Мода	=МОДА([x_i])
Размах вариации	=МАКС([x_i])-МИН([x_i])
Дисперсия	=ДИСПР([x_i])
Стандартное отклонение	=КОРЕНЬ(ДИСПР([x_i]))
Коэффициент вариации	= КОРЕНЬ(ДИСПР([x_i]))/СРЗНАЧ([x_i]) * 100
i-ый квартиль	=КВАРТИЛЬ([x_i]; i)
i-ый дециль	=ПЕРСЕНТИЛЬ([x_i]; i/10)
IQR	=КВАРТИЛЬ([x_i]; 3) - =КВАРТИЛЬ([x_i]; 1)
Нижняя граница для правила 3 сигм	=СРЗНАЧ([x_i]) – 3 * КОРЕНЬ(ДИСПР([x_i]))
Верхняя граница для правила 3 сигм	=СРЗНАЧ([x_i]) + 3 * КОРЕНЬ(ДИСПР([x_i]))

Результатов получилось достаточно много, поэтому полностью с ними можно ознакомиться в Приложении 1. Здесь же я лишь проинтерпретирую основные получившиеся значения.

Мода – наиболее часто встречающееся в числовом ряду значение. Мы видим отсутствие моды для признаков №3 и №5. Это связано с большим диапазоном значений этих признаков: они имеют самый большой размах вариации и самую большую дисперсию по сравнению с остальными.

Коэффициент вариации показывает степень изменчивости по отношению к среднему показателю выборки. В нашем случае для всех признаков кроме признака №4 значение этого коэффициента превышает 10%, что говорит о высокой колеблемости и неоднородности статистической совокупности. Тем временем для

признака №4 данный показатель очень мал, что связано с маленьким значением стандартного отклонения. Значит субъекты РФ в 2020 году имели схожий индекс потребительских цен.

Квантили - это точки среза, делящие наблюдения в выборке на непрерывные интервалы с равными вероятностями. Например, первый квантиль для признака №5 (числа студентов) равен 8093. Это означает, что в 25% регионов число студентов очной формы обучения меньше или равно этому числу. Второй квартиль покажет меньше какого числа количество студентов в 50% регионов и так далее. Аналогично и с децилями, однако в них шаг вместо 25% будет составлять 10%.

Теперь рассмотрим интерквартильный размах, или иначе IQR. Он представляет собой порядковую статистику, численно равную разности между 1-м и 3-м квартилями распределения. Можно так же сказать, что интерквартильный размах это половина выборки, центрированная относительно медианы. Данный показатель в дальнейшем будет использован для обнаружения выбросов.

Z-преобразование

Z-преобразование (или стандартизация) - это процесс преобразования исходных данных в стандартный вид, который позволяет нам лучше понимать и анализировать данные. В статистике Z-преобразование часто используется для преобразования данных в нормальное распределение, что может быть полезно при применении некоторых статистических методов.

Суть Z-преобразования заключается в том, чтобы вычесть среднее значение из каждого значения в наборе данных и затем поделить полученную разность на стандартное отклонение. Формально, для каждого значения x в наборе данных Z-преобразование вычисляется по формуле:

$Z = \frac{x-\bar{x}}{\sigma}$, где $\bar{x}$ - среднее, σ – стандартное отклонение

С данными, которые получились после данного преобразования можно ознакомиться на листе “Z-преобразование” расчетного файла Excel.

Правило трёх сигм

Правило трех сигм заключается в том, что при нормальном распределении практически все значения случайной величины с вероятностью 0,9973 лежат не далее трех сигм в любую сторону от математического ожидания, то есть находятся в диапазоне [μ−3σ; μ+3σ]. В нашем случае вместо математического ожидания используется среднее арифметическое, так как мы рассматриваем не выборку, а генеральную совокупность. Сигмой же обозначается стандартное отклонение. Мы предполагаем, что значения каждого из пяти исследуемых признаков распределены нормально, а поэтому можем воспользоваться данным правилом.

Подставив нужные значения в формулу получаем следующие диапазоны:

Таблица 2. Диапазоны значений правила трёх сигм для каждого из признаков

Признак 1	[-5.234023543, 19.26931766]
Признак 2	[-1.787162875, 3.076104052]
Признак 3	[-8.343392, 22.517325]

Признак 4	[102.521269, 107.781084]
Признак 5	[-149256.5171, 207486.5877]

Заметим, что нижние границы значений у некоторых из признаков получились отрицательными, что невозможно так как все признаки кроме третьего по смыслу больше нуля или равны ему. Поэтому исправим интервалы для этих признаков, поставив в качестве левой границы минимальное возможное значение – нуль. Также округлим значения до двух знаков после запятой. Получим следующие диапазоны:

Таблица 3. Исправленные диапазоны значений правила трёх сигм

Признак 1	[0, 19.27]
Признак 2	[0, 3.08]
Признак 3	[-8.34, 22.52]
Признак 4	[102.52, 107.78]
Признак 5	[0, 207486.59]

Это означает, что около 99,7% значений каждой из переменных лежат в соответствующем диапазоне.

Правило 1.5 и 3 IQR

Правило 1.5 и 3 IQR - это статистические методы для определения выбросов в наборе данных.

Правило 1.5 IQR гласит, что любое значение, находящееся вне диапазона от Q1-1.5*IQR до Q3 + 1.5*IQR, считается выбросом. Правило 3 IQR аналогично, но использует множитель 3 вместо 1.5, что приводит к более жестким критериям выбросов.

Воспользовавшись данными правилами, а также убрав отрицательные значения аналогично тому, как я делала для правила трех сигм, получаем следующие границы:

Таблица 4. Границы по правилу 1.5 IQR

Признак 1	Признак 2	Признак 3	Признак 4	Признак 5
[0.7, 11.9]	[0, 1.735]	[-5.29, 18.89]	[103.22, 107.22]	[0, 69575.5]

Таблица 5. Границы по правилу 3 IQR

Признак 1	Признак 2	Признак 3	Признак 4	Признак 5
[0, 13,3]	[0, 2.05]	[-14.36, 21.91]	[101.72, 107.72]	[0,81872]

Диаграммы

Теперь, чтобы получить еще более полное представление о наших данных, визуализируем их, построив для них точечные, листовые и ящичковые диаграммы.

Dotplot (точечная диаграмма) - это графическое представление данных, которое используется для визуализации распределения значений непрерывной переменной. Она состоит из одиночных точек, расположенных на оси координат, которые

показывают распределение значений переменной. С точечными диаграммами для каждой из пяти переменных можно ознакомиться ниже на Рис.1 – Рис.5.

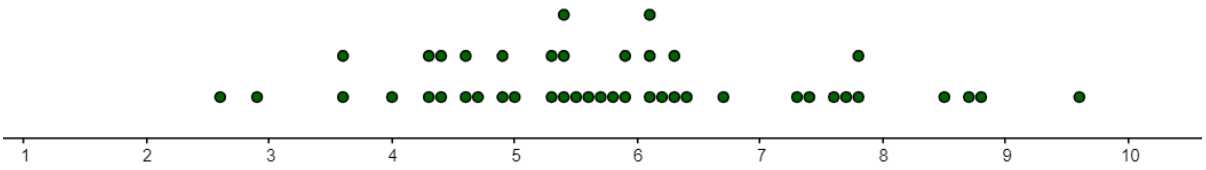

Рис 1. Точечная диаграмма для признака 1

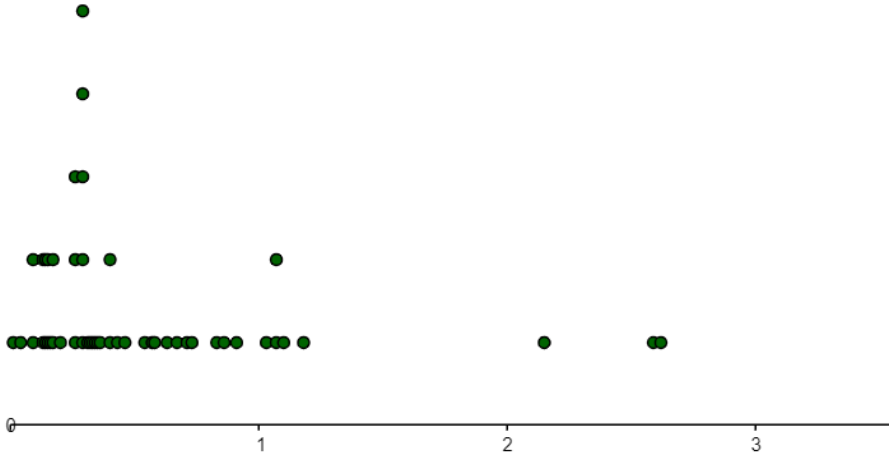

Рис.2 Точечная диаграмма для признака 2

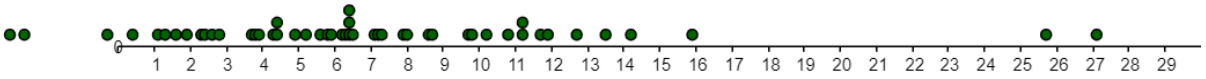

Рис.3 Точечная диаграмма для признака 3

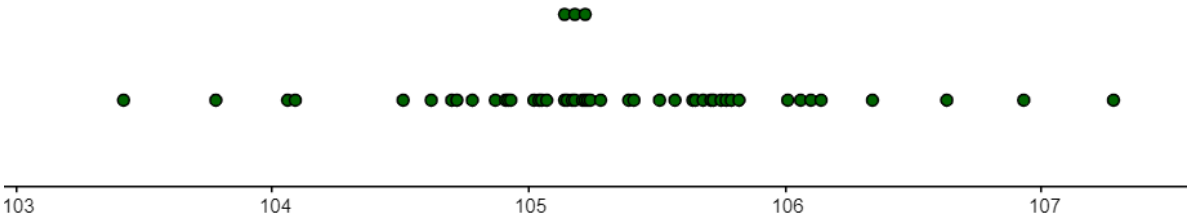

Рис.4 Точечная диаграмма для признака 4

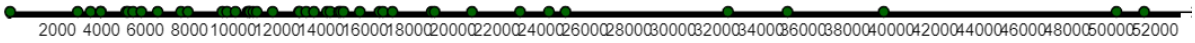

Рис.5 Точечная диаграмма для признака 5

Из-за большого размаха вариации признака 5 изобразить точечную диаграмму для нее очень проблематично. Более того, эта диаграмма будет мало информативна, в чем можно убедиться, посмотрев на Рис.5.

Ящичковые диаграммы позволяют визуализировать и сравнивать распределение и основную тенденцию числовых значений посредством их квартилей. Также они позволяют находить выбросы в данных. Рассмотрим по очереди этот вид диаграммы для каждого из признаков.

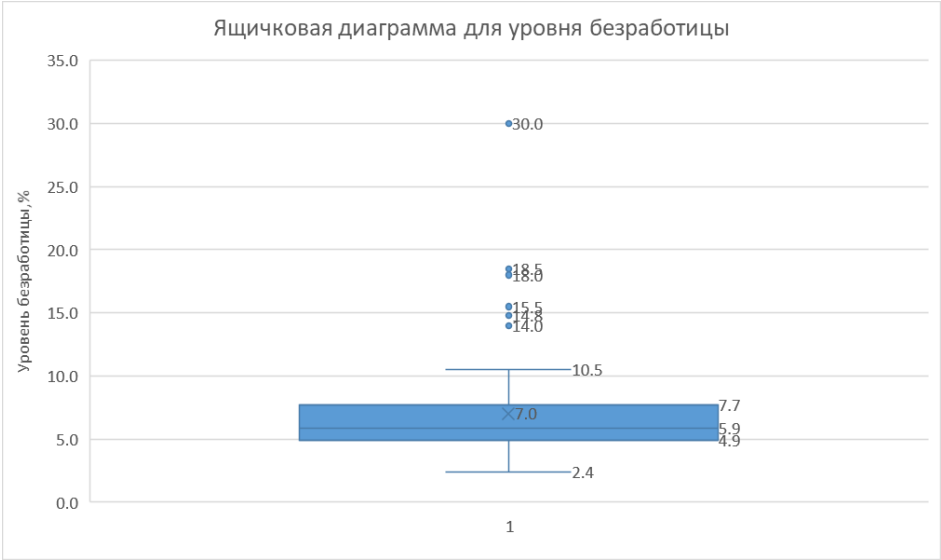

Рис.6 Ящичковая диаграмма для уровня безработицы

На Рис.6 видно, что 50% регионов имеют уровень безработицы от 4.9 до 7.7 %, о чем нам говорит расположение синего прямоугольника. Также на нем мы можен наблюдать значения квантилей и среднего арифметического, которые совпадают с уже посчитанными нами ранее знаечниям. Стоит отметить выбросы – это точки расположенные отдельно от основной фигуры на графике. В данном случае их 6 и по исходным данным несложно определить, что это такие субъекты РФ как Республика Ингушетия (30%), Чеченская Республика (18.5%), Республика Тыва (18%), Республика Северная Осетия–Алания(15.5%), Кабардино-Балкарская Республика (14.8%), Карачаево-Черкесская Республика(14.8%) и Республика Алтай (14%). Очевидно, что такой высокий уровень безработицы в данных субъектах вызван их расположением в горных районах, а также популярностью земледелия среди местных жителей.

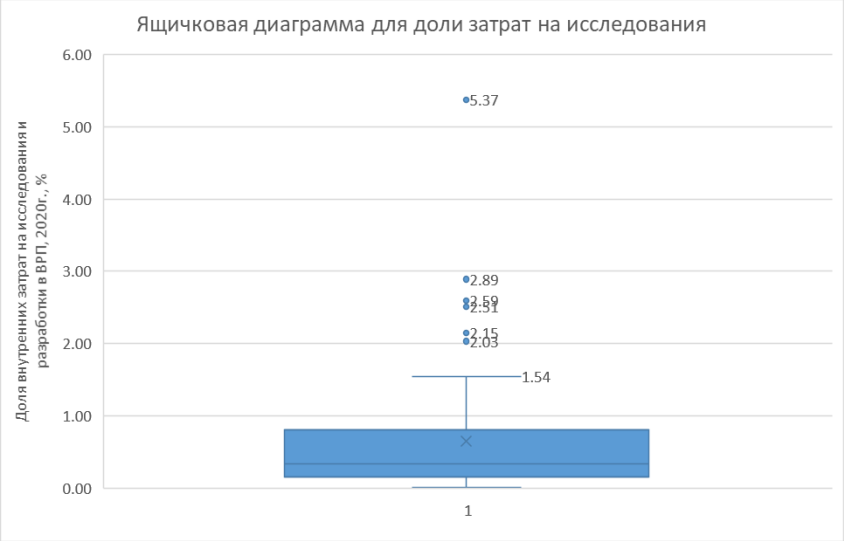

Рис.7 Ящичковая диаграмма для доли внутренних затрат на исследования и разработки

На Рис. 7 представлена ящичковая диаграмма уже для второго признака. Здесь также имеются выбросы. Их 6 штук и это такие регионы как Нижегородская область (5.37%), Томская область (2.89%), г. Санкт-Петербург (2.59%), Ульяновская область (2.51%), г.Москва (2.15%) и Новосибирская область (2.03%). Объяснить это можно либо тем, что данные субъекты являются крупнейшими городами страны с большим количеством научно-исследовательских центров (в случае Москвы и Санкт-Петербурга), либо тем, что в этих регионах в советское время были построены закрытые города с научными институтами, где велись различные разработки. Многие из них до сих пор продолжают функционировать.

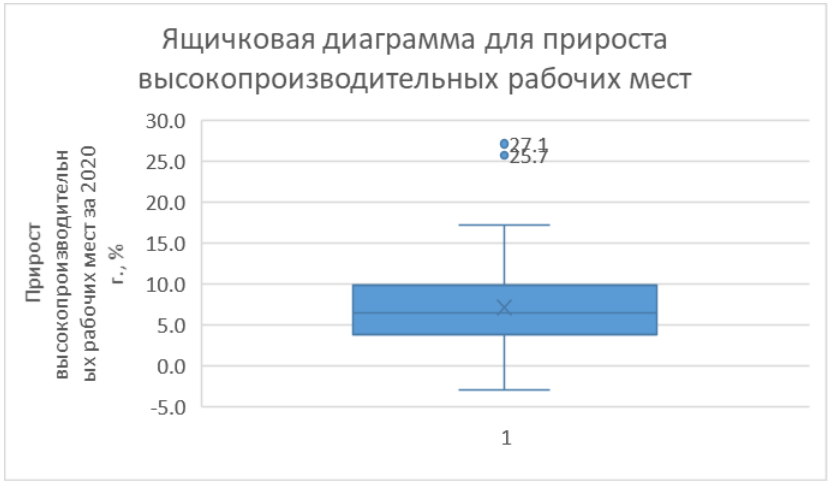

Рис.8 Ящичковая диаграмма для прироста высокопроизводительных рабочих мест

Судя по Рис.8 наибольший прирост высокопроизводительных рабочих мест наблюдается в Республике Калмыкия (25.7%) и в Республике Ингушетия (27.1%).

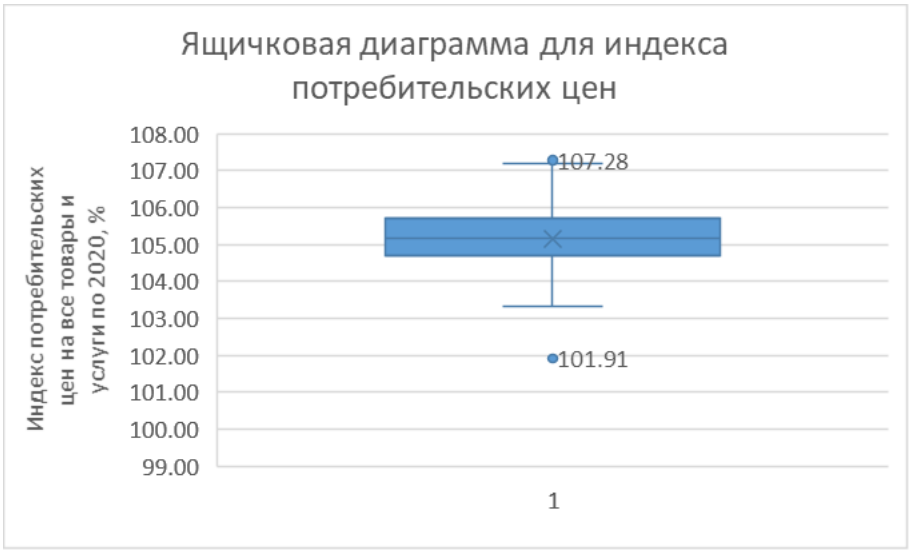

Рис.9 Ящичковая диаграмма для индекса потребительских цен

По Рис.9 видно, что выбросами с точки зрения индекса потребительских цен являются такие регионы как Республика Дагестан (107.28%) и Чукотский автономный округ (101.91%).

Рис.10 Ящичковая диаграмма для численности студентов очной формы обучения

Исходя из Рис.10 сильно выделяется количество студентов в городе Москва (500339 чел.), г.Санкт-Петербург (236664 чел.), Республике Татарстан (90945 чел.) и Свердловской области (72897 чел.). Такое большое количество обучающихся связано с большим количеством населения в данных регионах.

Мной также были построены листовые диаграммы для исследуемых признаков. В связи с неудобством их отображения в Word с ними можно ознакомиться на листе “Предварительный анализ” расчетного файла Excel.

Выводы

В данном разделе были посчитаны основные статистические показатели для каждой из пяти переменных. Выяснилось, что наибольший размах наблюдается по переменной №5 (численности студентов очной формы обучения), а наименьший – по переменным №2 и №4 (индексу потребительских цен и доле внутренних затрат на исследования). В целом стало понятно, что исходные данные достаточно неоднородны и имеют множество выбросов. Были построены ящичковые диаграммы, которые помогли эти выбросы обнаружить.

Корреляционный анализ

Начнем наш корреляционный анализ с построения полей корреляции. Поле корреляции - это графическое изображение значений коэффициента корреляции между двумя переменными для каждой пары значений этих переменных. Оно используется для визуализации силы и направления связи между двумя переменными в исследуемом наборе данных.

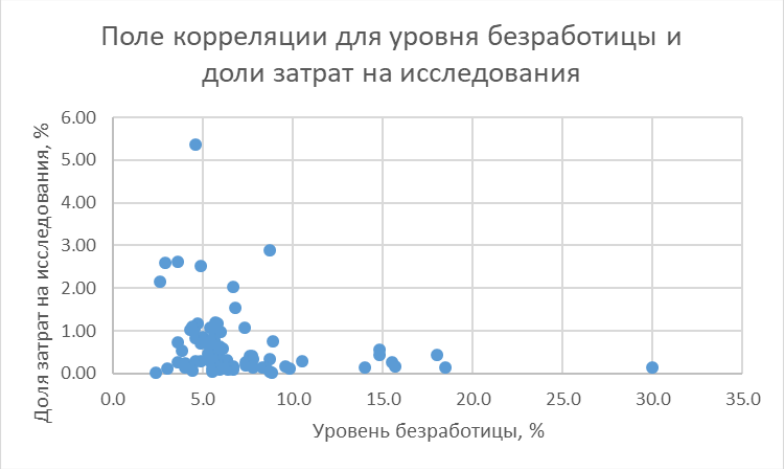

Рис. 11 Поле корреляции для признаков 1 и 2

Как видно из данной диаграммы, у регионов с небольшим уровнем безработицы доля затрат на исследования в среднем выше, чем у регионов с высоким уровнем безработицы. Этот результат вполне ожидаемый, ведь высокий уровень безработицы зачастую говорит о неблагополучии региона, а в таком регионе вряд ли найдутся средства на финансирование науки и исследований. Поэкспериментировав с разными линиями тренда в Excel, я выяснила, что лучше всего данный график приближается полиномиальной функцией.

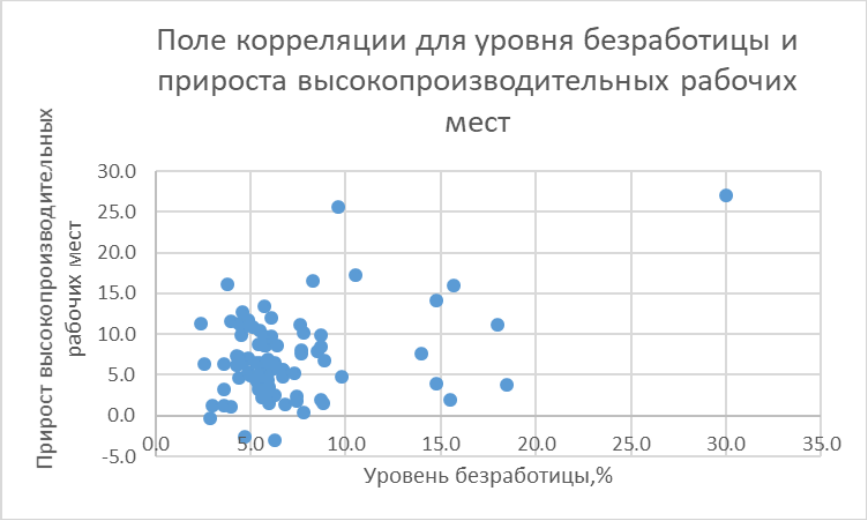

Рис.12 Поле корреляции для признаков 1 и 3

Диаграмма на Рис.12 показывает корреляцию между уровнем безработицы в регионе и приростом высокопроизводительных рабочих мест. Результаты несколько контринтуитивны: чем больше уровень безработицы в регионе, тем в среднем быстрее растет число высокопроизводительных рабочих мест. Связано это, вероятно, с тем, что государство предпринимает меры по борьбе с безработицей в таких регионах и пытается создавать там новые рабочие места. В то время как в регионах с низким уровнем безработицы не возникает потребности в увеличении числа рабочих мест.

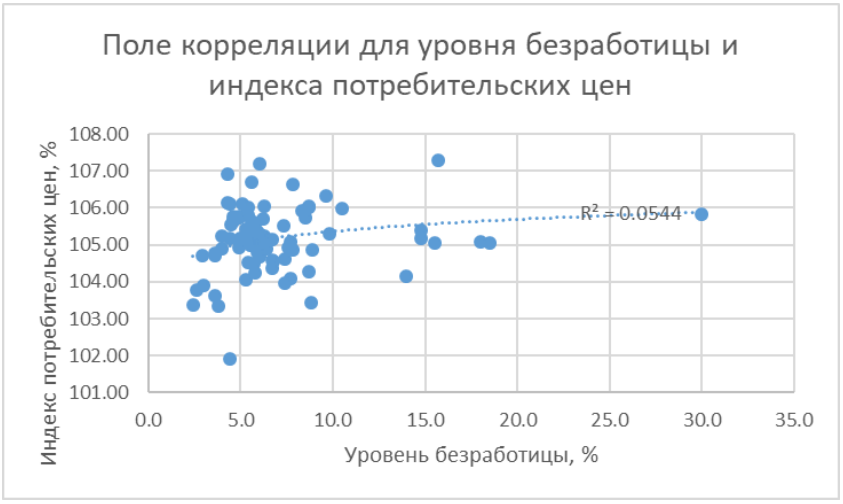

Рис. 13 Поле корреляции для признаков 1 и 4

Поле корреляции между уровнем безработицы и индексом потребительских цен изображено на Рис. 13 и лучше всего приближается логарифмической функцией. График не обладает четкой структурой и по нему сложно выявить какие-либо сильные взаимосвязи, однако все же наблюдается небольшой тренд на рост индекса потребительских цен по мере роста уровня безработицы.

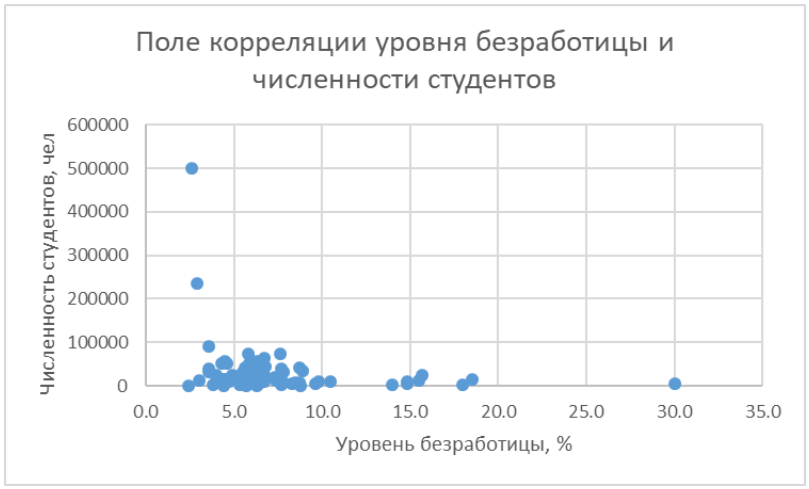

Рис. 14 Поле корреляции для признаков 1 и 5

На Рис. 14 показана корреляция между уровнем безработицы в регионе и численностью студентов очной формы обучения. На графике явно заметно экспоненциальное расположение точек. Чем ниже уровень безработицы, тем выше численность студентов и наоборот. В регионах с высокой безработицей число студентов очень близко к нулю.

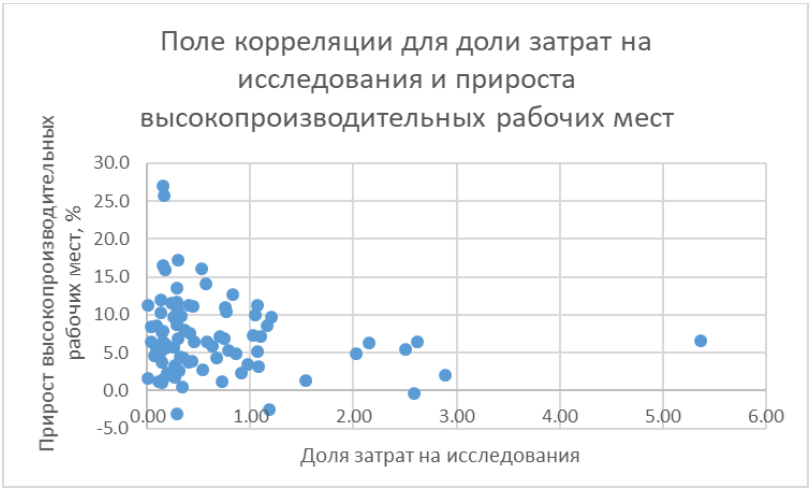

Рис. 15 Поле корреляции для признаков 2 и 3

На Рис.15 изображена корреляция между долей затрат на исследования и приростом высокопроизводительных рабочих мест. Явной взаимосвязи между признаками не наблюдается, можно отметить разве что очень слабую обратную взаимосвязь. Чем больше доля затрат на исследования, тем чуть меньше становится прирост высокопроизводительных рабочих мест. Однако эта взаимосвязь слишком слабая, чтобы делать какие-либо выводы.

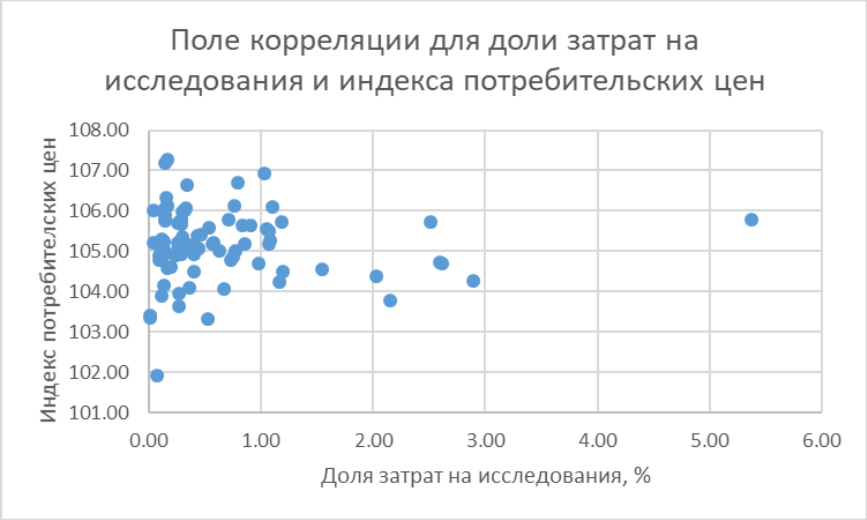

Рис. 16 Поле корреляции для признаков 2 и 4

На данной диаграмме изображена корреляция между долей затрат на исследования и индексом потребительских цен. По графику не видно никакой зависимости между данными признаками.

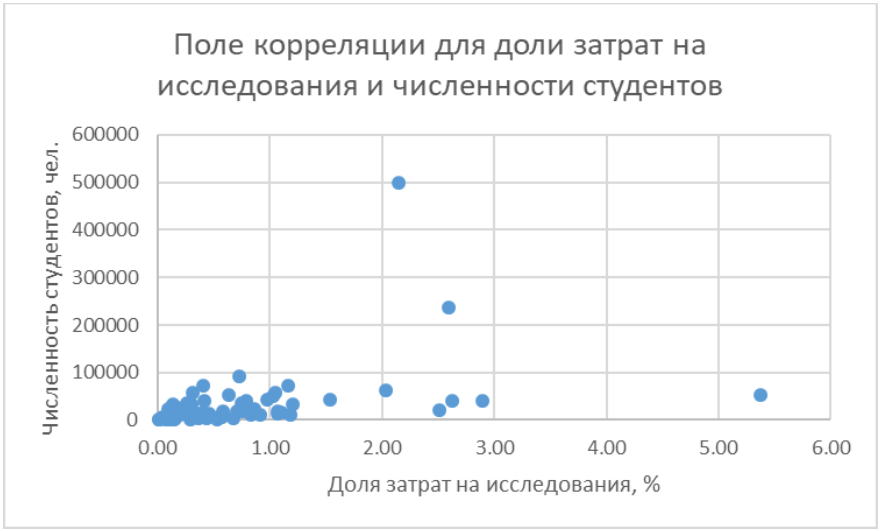

Рис. 17 Поле корреляции для признаков 2 и 5

На Рис.17 можно увидеть корреляцию между долей затрат на исследования и численностью студентов очной формы обучения по регионам РФ. Можно отметить некоторую положительную зависимость: чем выше доля затрат на исследования, тем чуть больше становится студентов. Причины данной взаимосвязи очевидны: если регион занимается развитием науки и финансирует исследования, то у большего числа людей есть мотивация поступать в высшие учебные заведения, чтобы в дальнейшем заниматься наукой.

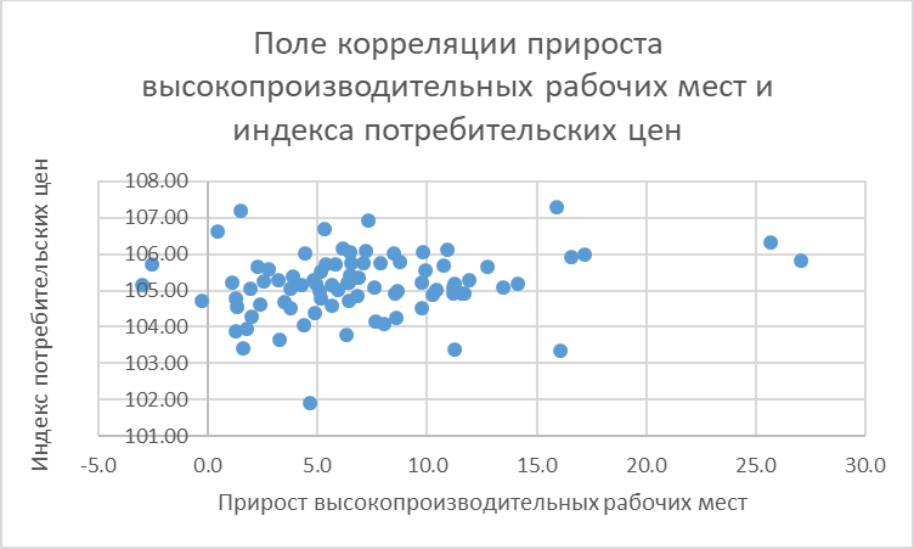

Рис. 18 Поле корреляции для признаков 3 и 4

На диаграмме выше изображена корреляция между приростом высокопроизводительных рабочих мест и индексом потребительских цен. Наблюдается слабая положительная взаимосвязь, т.е. с ростом числа рабочих мест немного растет индекс потребительских цен.

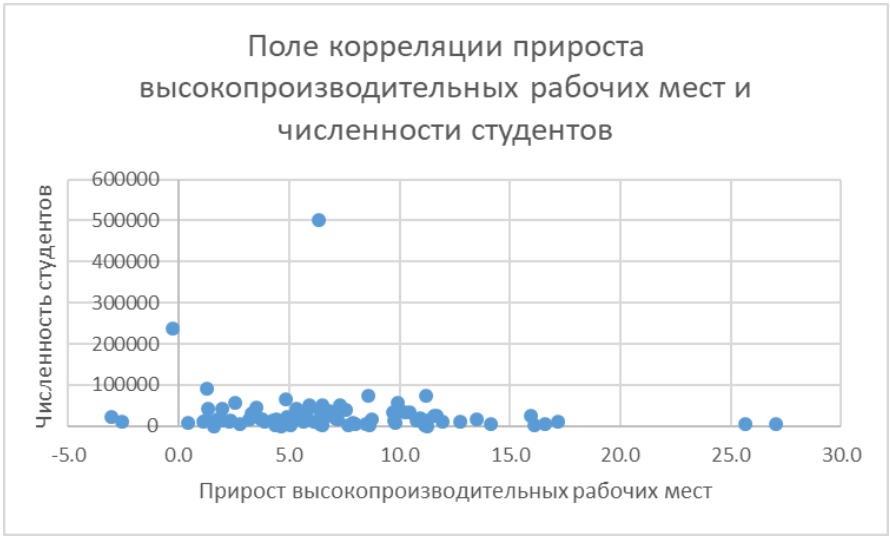

Рис. 19 Поле корреляции для признаков 3 и 5

На Рис.19 мы видим поле корреляции прироста высокопроизводительных рабочих мет и численности студентов по регионам РФ. Никакой связи между исследуемыми признаками по графику не наблюдается.

Рис. 20 Поле корреляции для признаков 4 и 5

На Рис.20 изображено поле корреляции для индекса потребительских цен и численности студентов. Никакой явной связи между исследуемыми признаками по графику не наблюдается.

Теперь построим матрицу парных коэффициентов корреляции с помощью пакета “Анализ данных” в Excel.

Корреляционная матрица представляет собой таблицу, на пересечении строк и столбцов которой находятся коэффициенты корреляции между соответствующими значениями. Коэффициент корреляции отражает степень взаимосвязи между двумя показателями. Всегда принимает значение от -1 до 1. Если коэффициент расположился около 0, то говорят об отсутствии связи между переменными. Если же значение коэффициента близко к 1, то имеется сильная взаимосвязь. Полученный результат отображен на Рис.21

	Признак 1	*Признак 2*	*Признак 3*	*Признак 4*	*Признак 5*
Признак 1	1				
Признак 2	-0.222560516	1			
Признак 3	0.379980554	-0.17048403	1		
Признак 4	0.17227078	-0.01700329	0.194661826	1	
Признак 5	-0.211753884	0.409388941	-0.122663857	-0.1719095	1

Рис.21 Матрица парных коэффициентов корреляции

Матрица парных коэффициентов корреляции подтверждает выводы о силе и характере связи, сделанные нами ранее по графикам. Например, для признаков 1 и 2 коэффициент корреляции равен -0.22. Отрицательное значение говорит об обратной взаимосвязи, а число 0.22, находящееся не так далеко от нуля, означает слабый характер связи. Для признаков 2 и 5 наоборот: значение 0.41 говорит о положительной взаимосвязи, гораздо более сильной чем между признаками 1 и 2. Аналогично интерпретируются и остальные значения из таблицы.

Теперь займемся удалением выбросов из наших данных. Какие именно регионы мы будем удалять мы уже знаем из первой части работы, где с помощью ящичковых диаграмм мы исследовали данные на выбросы. Этими регионами станут: Республика Ингушетия, Чеченская Республика, Республика Тыва, Республика Северная Осетия–Алания, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Дагестан, Республика Алтай, Республика Татарстан, Нижегородская область, Томская область, г. Москва и г.Санкт-Петербург. Данные с убранными выбросами представлены на листе “Данные без выбросов” расчетного файла Excel.

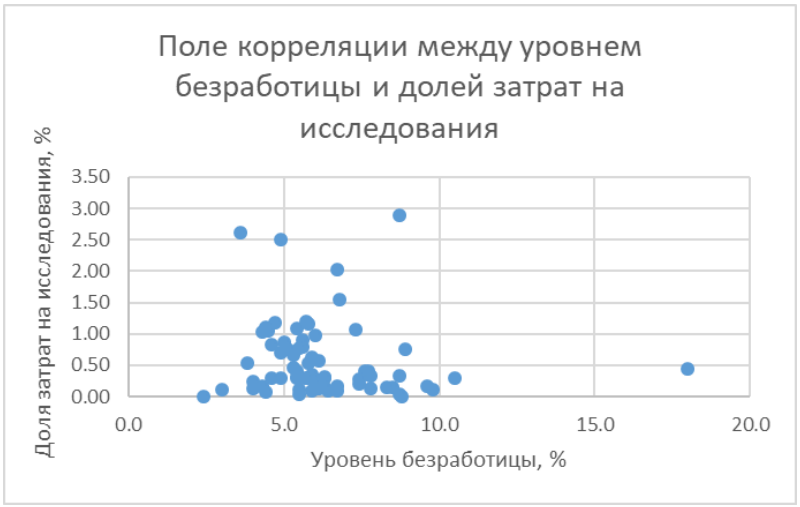

Рис. 22 Поле корреляции между признаками 1 и 2 после удаления выбросов

Не будем перестраивать все поля корреляции, так как очевидно, что они все претерпят примерно одинаковые изменения. Рассмотрим их на примере корреляции между признаками 1 и 2, изображенной на Рис.12. Сравнив данный рисунок с Рис. 6 заметим, что исчезли точки, располагавшиеся вдали от остальных. Другая же часть графике не претерпела изменений.

Возможно на некоторых из них чуть заметнее станет взаимосвязь между переменными, но этот эффект мы сможем пронаблюдать и построив новую матрицу парных коэффициентов корреляции.

	Признак 1	*Признак 2*	*Признак 3*	*Признак 4*	*Признак 5*
Признак 1	1				
Признак 2	-0.1218905	1			
Признак 3	0.1630487	-0.16463371	1		
Признак 4	0.101925	-0.01869736	0.10603981	1	
Признак 5	-0.0841419	0.46997176	-0.07543575	-0.0031287	1

Рис.23 Матрица парных коэффициентов корреляции (после удаления выбросов)

По Рис.23 мы наблюдаем следующий эффект от удаления выбросов: большинство коэффициентов корреляции (за исключением коэффициента для

признаков 2 и 5) по модулю приблизились к нулю. Это значит, что стала прослеживаться еще меньшая по силе взаимосвязь между признаками. Причиной этого стал тот факт, что по каждой переменной у нас имеется большой количество выбросов и когда мы их всех удалили наш датасет сильно уменьшился в размерах. Теперь он составляет 73 наблюдения вместо исходных 86, что меньше на целых 16%. Чем меньше у нас становится данных, тем сложнее выявить наличие, силу и характер взаимосвязи между переменными.

Выводы о связи между признаками

Все изучаемые нами признаки, кроме пары из 2 и 5, довольно слабо взаимосвязаны друг с другом. Между некоторыми призаками (1 и 3, 1 и 4, 2 и 5, 3 и 4) наблюдается положительная связь, между остальными парами – отрицательная. Довольно сильно коррелируют признаки 2 и 5 и почти совсем не коррелируют признаки 2 и 4, 4 и 5.

Частные коэффициенты корреляции

Теперь будем строить матрицу частных коэффициентов корреляции. Частный коэффициент корреляции характеризует тесноту линейной зависимости между результатом и соответствующим фактором при устранении влияния других факторов. Частный коэффициент корреляции оценивает тесноту связи между двумя переменными при фиксированном значении остальных факторов. Если вычисляется, например, $r_{yx|z}$ (частный коэффициент корреляции между y и x при фиксированном влиянии z), это означает, что определяется количественная мера линейной зависимости между y и x, которая будет иметь место, если устранить влияние z на эти признаки.

Частные коэффициенты корреляции первого порядка вычисляются по формуле: $r_{ij,s} = -\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}}$ где R_{ij} - алгебраическое дополнение элемента r_{ij} матрицы R.

$$r_{x1x2/x3} = \frac{-0.223 - 0.38 * (-0.171)}{\sqrt{(1 - 0.38^2)(1 - 0.171^2)}} = -0.173$$

Теснота связи низкая. Аналогично посчитаем и для всех остальных троек коэффициентов. Получим следующую таблицу.

r 12|3	-0.173	r 15|2	-0.136	r 25|1	0.38	r 45|1	-0.412
r 12|4	-0.223	r 15|3	-0.18	r 25|3	0.397	r 45|2	-0.181
r 12|5	-0.152	r 15|4	-0.188	r 25|4	0.413	r 45|3	-0.152
r 13|2	0.356	r 23|1	-0.095	r 34|1	0.141		
r 13|4	0.359	r 23|4	-0.171	r 34|2	0.194		
r 13|5	0.365	r 23|5	-0.133	r 34|5	0.177		
r 14|2	0.173	r 24|1	0.022	r 35|1	-0.047		
r 14|3	0.108	r 24|3	0.016	r 35|2	-0.059		
r 14|5	0.141	r 24|5	0.059	r 35|4	-0.093		

Рис. 24 Таблица со значениями выборочных частых коэффициентов корреляции первого порядка

Теперь мы хотим получить частные коэффициенты корреляции более высоких порядков, то есть те, при которых фиксируется теснота связи двух переменных при устранении влияния двух и более факторов. Коэффициенты частной корреляции более высоких порядков можно найти через коэффициенты частной корреляции более низких порядков по рекуррентной формуле:

$$r_{x_1x_2|x_3\dots x_n} = \frac{r_{x_1x_2|x_3\dots x_{n-1}} - r_{x_1x_n|x_3\dots x_{n-1}} \times r_{x_2x_n|x_3\dots x_{n-1}}}{\sqrt{(1 - r_{x_1x_n|x_3\dots x_{n-1}}{}^2)(1 - r_{x_2x_n|x_3\dots x_{n-1}}{}^2)}}$$

Полученные значения коэффициентов частной корреляции второго порядка представлены на Рис.245

r 12|34	-0.176	r 15|23	-0.123	r 25|13	0.378	r 45|12	-0.455
r 12|35	-0.092	r 15|24	-0.108	r 25|14	0.427	r 45|13	-0.410
r 12|45	-0.162	r 15|34	-0.166	r 25|34	0.404	r 45|23	-0.173
r 13|24	0.334	r 23|14	-0.099	r 34|12	0.144		
r 13|25	0.352	r 23|15	-0.083	r 34|15	0.134		
r 13|45	0.349	r 23|45	-0.146	r 34|25	0.187		
r 14|23	0.113	r 24|13	0.036	r 35|12	-0.012		
r 14|25	0.152	r 24|15	0.212	r 35|14	0.012		
r 14|35	0.083	r 24|35	0.084	r 35|24	-0.025		

Рис. 25 Коэффициенты частной корреляции второго порядка

Теперь наконец можно приступить к подсчету коэффициентов третьего порядка. После произведения всех расчетов соберем их в матрицу, аналогичную матрице парных коэффициентов корреляции, чтобы произвести сравнение. Полученная матрица изображена ниже на Рис.26

	Признак 1	Признак 2	Признак 3	Признак 4	Признак 5
Признак 1	1.000				
Признак 2	-0.120	1.000			
Признак 3	0.333	-0.115	1.000		
Признак 4	0.094	0.226	0.155	1.000	
Признак 5	-0.106	0.430	0.061	-0.458	1.000

Рис. 26 Матрица коэффициентов частной корреляции третьего порядка

Сравним полученную матрицу с матрицей, изображенной на Рис. 21. Сильно увеличилась корреляция между признаками 2 и 4 (была -0.017, стала 0.226), более того значение коэффициента корреляции изменило знак, то есть поменялось в том числе направление зависимости. Значит с ростом доли внутренних затрат на исследования индекс потребительских цен тоже растет. Стоит также отметить корреляцию между признаками 3 и 5. Изначально она была равна -0.122, но после расчетов ее значение стало 0.061. То есть знак также изменился. Сильно выросла по модулю корреляция между признаками 4 и 5. Она составляла -0.171, а по итогу стала -0.458. Это значение говорит о том, что взаимосвязь между данными признаками достаточно значительная.

Далее хотим посчитать множественный коэффициент корреляции для переменной, которая будет использована в регрессионном анализе в качестве зависимой переменной. В нашем случае это будет переменная №5, т.е. число студентов очной формы обучения по регионам РФ. Сделаем это по следующей формуле:

$R_{x_5|x_1x_2x_3x_4} = \sqrt{\frac{\Delta_{yx}}{\Delta_{xx}}}$, где Δ_{yx}- определитель всей матрицы корреляции, а Δ_{yx} – алгебраическое дополнение элемента r_{55} корреляционной матрицы.

Получаем значение $R_{x_5|x_1x_2x_3x_4} = \sqrt{\frac{\Delta_{yx}}{\Delta_{xx}}} = \sqrt{\frac{0.347}{0.793}} = 0.661$

Множественный коэффициент корреляции (также называемый коэффициентом детерминации) - это мера степени связи между зависимой переменной и набором независимых переменных в множественной регрессии. Коэффициент детерминации принимает значения от 0 до 1 и показывает долю дисперсии зависимой переменной, объясненную набором независимых переменных.

Значение множественного коэффициента корреляции равное 0.661 означает, что 66,1% дисперсии зависимой переменной можно объяснить набором независимых переменных в модели множественной регрессии. Оставшиеся 33,9% дисперсии могут быть объяснены другими факторами, которые не были учтены в модели.

Выводы

После проведенного корреляционного анализа можно сделать следующие выводы:

1. Корреляция между каждой парой изучаемых признаков достаточно небольшая. Ни одно значение парного коэффициента корреляции не превышает по модулю 0.5, что говорит о слабом характере связи между признаками.
2. Удаление выбросов не помогает получить более точное представление о корреляции между величинами, так как выбросов много и при их удалении мы теряем большое количество информации
3. Построение частных коэффициентов корреляции позволяет исключить влияние других переменных на связь между двумя изучаемыми переменными, и показать, как сильно эти две переменные связаны, учитывая влияние всех остальных факторов.
4. Наша модель объясняет более чем половину изменения зависимой переменной, что может быть рассмотрено как хороший показатель качества модели.

Кластерный анализ данных

Кластерный анализ данных — это метод в статистике, который позволяет группировать объекты в наборе данных на основе их сходства или расстояния между ними. Он используется для исследования структуры данных и выявления групп или кластеров объектов, которые имеют схожие характеристики или свойства.

Начнем наш кластерный анализ с выбора меры расстояния, которая будет использоваться для проведения иерархической кластеризации. Возможными вариантами являются: евклидово расстояние, расстояние Махаланобиса, манхэттенское расстояние и другие. Я буду использовать самое обычное евклидово расстояние. Евклидово расстояние - это самая распространенная мера расстояния, которая определяется как квадратный корень из суммы квадратов разностей между соответствующими координатами объектов. Она хорошо подходит для данных с нормальным распределением, где значения переменных не сильно отличаются друг от друга. Это как раз подходит для наших данных, так как мы считаем, что они распределены нормально. Можно было бы также использовать расстояние Махаланобиса, которое учитывает коррелированность случайных величин, но, как мы уже считали выше, наши признаки не сильно коррелированы, а значит особой разницы в результатах не будет.

Теперь поговорим про алгоритмы кластеризации. Их существует множество:

- Метод ближнего соседа (Single Linkage). Этот метод определяет расстояние между кластерами, используя расстояние между

ближайшими объектами из каждого кластера. Недостатком этого метода является его чувствительность к выбросам и шумам.

- Метод дальнего соседа (Complete Linkage). Этот метод определяет расстояние между кластерами, используя расстояние между самыми дальними объектами из каждого кластера. Он более устойчив к шуму, чем метод ближнего соседа, но может привести к проблеме "склеивания" кластеров.
- Центроидный метод (Centroid Linkage). Этот метод определяет расстояние между кластерами, используя расстояние между центроидам
- Метод средней связи (Average Linkage). Этот метод определяет расстояние между кластерами, используя среднее расстояние между всеми парами объектов в кластерах. Он более устойчив к выбросам, чем метод ближнего соседа, и менее склонен к "склеиванию" кластеров, чем метод дальнего соседа.
- Алгоритм К-средних (K-means algorithm): Этот алгоритм разбивает объекты на K кластеров, минимизируя сумму квадратов расстояний между объектами и центроидами их кластеров.

Кластеризуем наши данные методом наименьшего соседа. Для этого сначала построим таблицу 85 × 85, в которой будут указаны евклидовы расстояния между объектами. Эта таблица представлена в расчетном файле Excel на листе “Кластерный анализ”. Далее найдем клетку в таблице, в которой находится минимальное значение евклидова расстояния (клетки на диагонали не учитываем). В нашем случае это будет клетка, отвечающая за расстояние между регионами №62 и №24. Это регионы: Ханты-Мансийский автономный округ и Калининградская область. Объединим их в один кластер. После этого пересчитаем расстояния в таблице с учетом того, что эти два субъекта теперь образуют один кластер. Продолжим делать так дальше, пока расстояния между объединяемыми кластерами не начнут становиться большим, так как большое расстояние говорит о том, что мы объединяем непохожие объекты. К сожалению, возможности программы Excel не позволяют проделать данные действия с таблицей нашего размера за адекватное время, так что остановимся на описании данного алгоритма.

Кластеризуем наши данные методом дальнего соседа. Алгоритм метода дальнего соседа работает следующим образом: начинаем с создания n кластеров, где каждый объект представляет отдельный кластер. Рассчитываем расстояние между всеми парами кластеров, используя максимальное расстояние между всеми парами объектов в разных кластерах. Объединяем два ближайших кластера, которые имеют максимальное расстояние между объектами в разных кластерах. Повторяем предыдущие шаги до тех пор, пока все объекты не будут объединены в единственный кластер. Аналогично методу ближнего соседа, возможности программы Excel не

позволяют проделать данные действия с таблицей нашего размера за адекватное время, так что остановимся на описании данного алгоритма.

Кластеризация методом k-средних

Кластеризуем наши данные воспользовавшись методом k-средних. Для этого воспользуемся языком программирования Python, в котором уже встроены все нужные нам библиотеки и функции. Все расчеты целиком будут приведены в файле “cluster_analysis.ipynb”.

Для начала экспортируем наши данные из Excel-файла. После этого воспользуемся методом локтя для нахождения оптимального количества кластеров. Идея метода локтя заключается в том, чтобы построить график, где на оси X отображается количество кластеров, а на оси Y - критерий качества кластеризации. Обычно используется концепция минимизации внутри кластерной суммы квадратов (WCSS), которая является мерой того, насколько хорошо объекты сгруппированы в кластеры. Затем мы анализируем график и ищем место, где кривая начинает "изгибаться" и похожа на локоть. Это место будет предполагаемым оптимальным числом кластеров. Идея состоит в том, чтобы выбрать количество кластеров таким образом, чтобы мы получили максимальное улучшение качества кластеризации, при минимальном увеличении числа кластеров.

Стоит заметить, что в ходе вычислений у меня получалось, что такие регионы как Москва и Санкт-Петербург образуют собственные кластеры, в то время как все остальные субъекты РФ собираются в другие кластеры в большом количестве. Для повышения точности вычислений и более корректной кластеризации мной было решено избавиться от этих двух регионов как от выбросов и не учитывать их в кластерном анализе.

С помощью метода локтя мы получили следующий график:

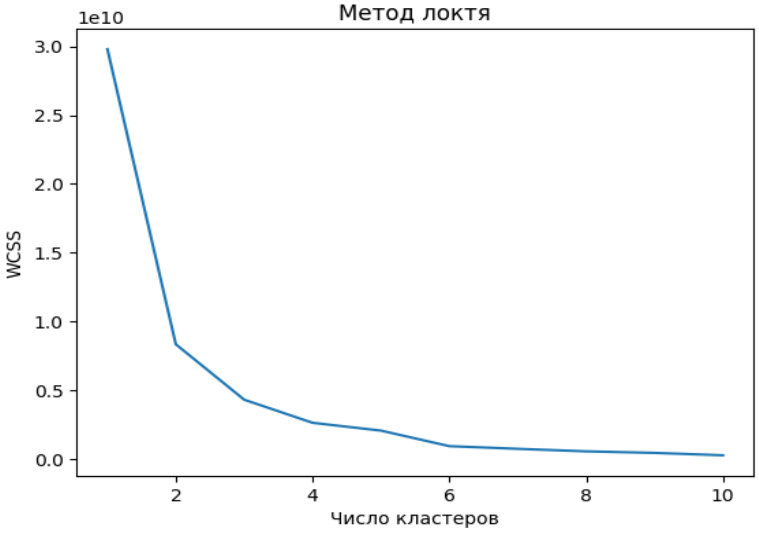

Рис. 27 График внутри кластерной суммы квадратов

Из Рис.27 видно, что внутри кластерная сумма квадратов начинает падать медленнее, если число кластеров начинает превышать 4. Значит мы можем использовать именно это значение при разбиении объектов на кластеры.

Нормализуем и стандартизируем данные с помощью модуля `sklearn.preprocessing`. Теперь воспользуемся функцией `KMeans` из библиотеки `sklearn.cluster` и к полученному результату применим функцию `fit_predict`. Получим вектор длины 85 с цифрами от 0 до 3. Это и есть номера кластеров, к которым был отнесен каждый из 85 регионов в ходе вычислений. Добавим в наш датасет новый столбец, в котором для каждого субъекта РФ будет указан полученный номер кластера.

Наши регионы разбились на 4 кластера следующим образом:

1) Белгородская область, Ивановская область, Курская область, Орловская область, Рязанская область, Смоленская область, Тамбовская область, Тверская область, Тульская область, Ярославская область, Калининградская область, Краснодарский край, Ростовская область, Республика Дагестан, Республика Северная Осетия - Алания, Чеченская Республика, Республика Мордовия, Удмуртская Республика, Чувашская Республика, Кировская область, Оренбургская область, Пензенская область, Ульяновская область, Ханты-Мансийский автономный округ - Югра, Кемеровская область, Республика Саха (Якутия), Хабаровский край
2) Воронежская область, Республика Крым, Волгоградская область, Республика Башкортостан, Республика Татарстан, Нижегородская область, Самарская область, Свердловская область, Новосибирская область
3) Брянская область, Владимирская область, Калужская область, Костромская область, Липецкая область, Республика Карелия, Республика Коми, Архангельская область, Ненецкий автономный округ, Вологодская область, Ленинградская область, Мурманская область, Новгородская область, Псковская область, Республика Адыгея, Республика Калмыкия, г. Севастополь, Республика Ингушетия, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Марий Эл, Курганская область, Ямало-Ненецкий автономный округ, Республика Алтай, Республика Тыва, Республика Хакасия, Республика Бурятия, Забайкальский край, Камчатский край, Амурская область, Магаданская область, Сахалинская область, Еврейская автономная область, Чукотский автономный округ
4) Московская область, Астраханская область, Ставропольский край, Пермский край, Саратовская область, Тюменская область, Челябинская область, Алтайский край, Красноярский край, Иркутская область, Омская область, Томская область, Приморский край

В каждом кластере получилось 27, 9, 34 и 13 регионов соответственно. Регионы распределились по кластерам достаточно равномерно, нет кластеров в которых было

бы 1 или 2 региона, а значит удаление Москвы и Санкт-Петербурга из рассмотрения было вполне достаточно.

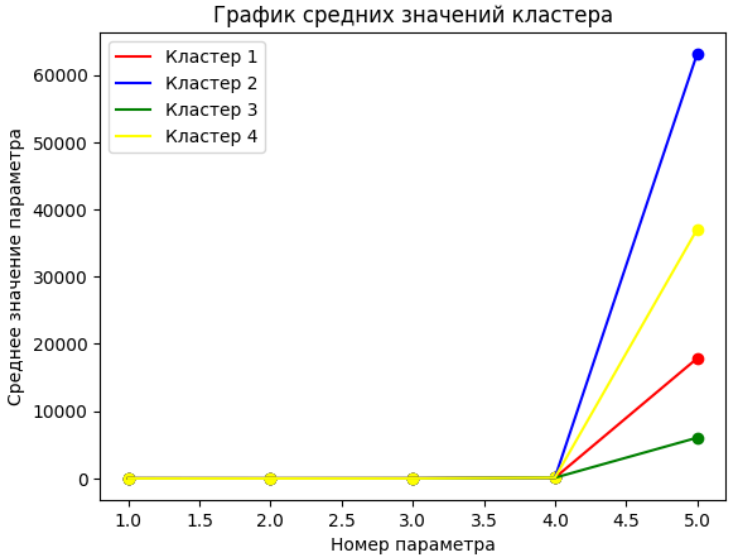

Рис.28 График средних значений кластера

На Рис. 28 приведен график средних значений кластера. На нем по оси абсцисс отмечается номер исследуемого признака, а по оси ординат – значение этого признака. На графике присутствуют линии четырех цветов, каждая из которых отвечает за значения в своем кластере. Из этого графика мы видим, что в основном регионы в кластерах отличаются по своему значению признака №5, то есть по количеству студентов. График выглядит именно так, потому что значения признака №5 самые большие по модулю. Однако мы хотим увидеть отличия и по другим признаками,для этого воспользуемся функцией plt.semilogy(), которая создаёт график с логарифмическим масштабированием по оси y. Получим следующий график:

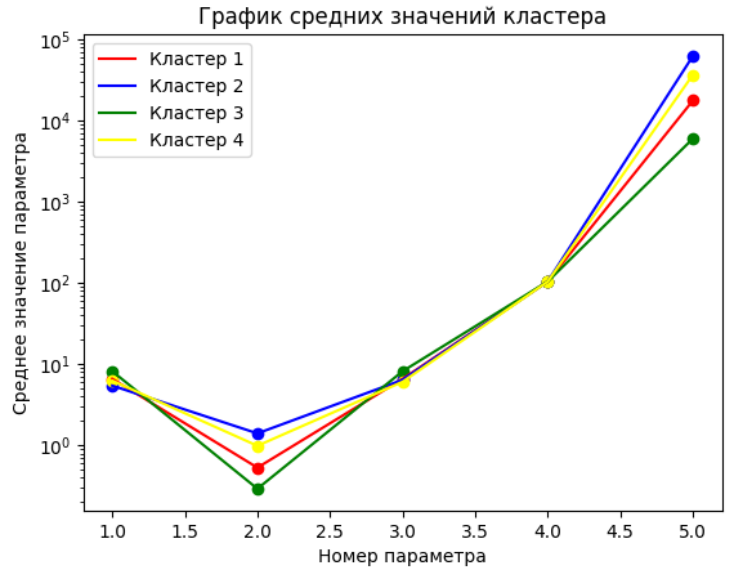

Рис. 29 График средних значений кластера на логарифмической шкале

Из измененного графика мы можем получить уже больше информации. Например, мы видим, что в кластере №3 находятся субъекты с самыми маленькими значениями признака 2 и самыми маленькими значениями признака 5. При этом средние значения признаков 1 и 3 у них наоборот больше, чем во всех других кластерах.

Получив новую информацию из графиков, перейдем к описанию кластеров.

Кластер №1: Скорей неблагополучные регионы

В этих субъектах РФ достаточно большая безработица (хоть и не самая большая в России), исследования финансируются редко, а людей, желающих получить очное образование немного.

Кластер №2: Благополучные регионы

В этих регионах почти нет проблем с безработицей, активно финансируется наука, много студентов очной формы обучения.

Кластер №3: Неблагополучные регионы

Здесь большой уровень безработицы, почти не выделяются деньги на исследования и разработки, а число студентов очной формы обучения чрезвычайно мало.

Кластер №4: Скорей благополучные регионы

В этих субъектах уровень безработицы хоть и не самый низкий по России, но вполне приемлемый. Исследования тут финансируются достаточно хорошо, а обучаться на очной форме обучения предпочитает достаточно большое количество человек.

Значения параметров №3 и №4 (прирост высокопроизводительных рабочих мест и индекс потребительских цен) очень мало отличаются между регионами. Именно поэтому сложно что-либо сказать по поводу значений этих признаков в каждом из кластеров.

Выводы

Воспользовавшись методом локтя, мы выяснили, что оптимальным числом кластеров для нашего исследования будет четыре. С помощью метода k-средних мы провели кластеризацию субъектов РФ и получили кластеры, определяющие экономическое положение региона. Регионы разделились на благополучные, скорее благополучные, скорее неблагополучные и неблагополучные.

Заключение

В данной работе были посчитаны и проанализированы различные статистические показатели для пяти различных показателей по регионам РФ за 2020 год. Был проведен предварительный, корреляционный и кластерный анализ данных. Были построены различные графики для визуального представления данных и их более удобной интерпретации. Изучаемые признаки были исследованы на зависимость друг от друга, регионы были поделены на кластеры.

Выводы по каждому отдельному шагу анализа указаны после вычисления каждого показателя в ходе работы. Если говорить о выводах в целом, из полученных результатов можно судить о том, что субъекты Российской Федерации делятся на 4 кластера исходя из их уровня благополучия: благополучные, скорее благополучные, скорее неблагополучные и неблагополучные. Это частично подтверждает гипотезу, сделанную во введении данной работы. Однако вычисления показали, что прирост высокопроизводительных рабочих мест и индекс потребительских цен настолько мало отличаются от региона к региону, что не почти не влияют на кластеризацию. По сути в данном исследовании их можно было не учитывать.

Такие регионы, как г.Москва и г.Санкт-Петербург являются выбросами в наших данных, поскольку число студентов очной формы в них сильно выделяется в большую сторону относительно других регионов. В случае кластерного анализа эти регионы следует не учитывать или хотя бы поместить каждый в свой отдельный кластер.

Приложения

Приложение 1

Значения основных статистических показателей для исследуемых пяти переменных

	Уровень безработицы	Затраты на исследования	Прирост рабочих мест	Индекс потребительских цен	Численность студентов
Среднее	7.0	0.6	7.1	105.2	29115.0
Медиана	5.9	0.3	6.4	105.2	14971.0
Мода	5.4	0.29	#Н/Д	105.72	#Н/Д
Размах вариации	27.6	5.4	30.1	5.4	500339.0
Коэффициент вариации	58.19457955	125.7690425	72.576224	0.83369094	204.2147074
Дисперсия	16.67815917	0.656982367	26.455107	0.768490381	3535156744
Стандартное отклонение	4.0838902	0.810544488	5.1434528	0.876635831	59457.18412
Q1	4.9	0.16	3.773066	104.72	8093
Q2	5.9	0.33	6.3973144	105.17	14971
Q3	7.7	0.79	9.8183876	105.72	32686
Q4	30	5.37	27.066199	107.28	500339
D1	4	0.11	1.5342375	104.072	2813
D2	4.6	0.15	3.1319604	104.576	5828.8
D3	5.3	0.208	4.4710651	104.876	9951.6
D4	5.56	0.29	5.3655191	105.032	12667.2
D5	5.9	0.33	6.3973144	105.17	14971
D6	6.24	0.448	7.2314254	105.274	18938.4
D7	7.2	0.726	8.6450063	105.626	24906.2
D8	8.34	0.99	10.803404	105.774	36162
D9	10.22	1.192	12.414197	106.084	51733.8
D10	30	5.37	27.066199	107.28	500339
IQR	2.8	0.63	6.0453216	1	24593
1.5 IQR	4.2	0.945	9.0679824	1.5	36889.5
3 IQR	8.4	1.89	18.135965	3	73779
срзнач - 3 * сигма	-5.234023543	-1.787162875	-8.343392	102.521269	-149256.5171
срзнач + 3 * сигма	19.26931766	3.076104052	22.517325	107.781084	207486.5877
Q1 - 1,5 IQR	0.7	-0.785	-5.294916	103.22	-28796.5
Q3 + 1,5 IQR	11.9	1.735	18.88637	107.22	69575.5
Q1 - 3 IQR	-3.5	-1.73	-14.3629	101.72	-65686
Q1 + 3 IQR	13.3	2.05	21.909031	107.72	81872