Национальный исследовательский университет «Высшая школа экономики»

Московский институт электроники и математики им. А.Н. Тихонова НИУ ВШЭ

Образовательная программа: Информатика и вычислительная техника

Модульная домашняя работа 1

по майнору «Прикладной статистический анализ»

Работу выполнил

Фаткиев Айдар Ришатович

Москва, 2024 г.Содержание

[Введение 2](#_Toc163052898)

[Предварительный анализ данных 3](#_Toc163052899)

[Корреляционный анализ: 11](#_Toc163052900)

[Кластерный анализ: 17](#_Toc163052901)

[Общие выводы по работе 23](#_Toc163052902)

[Используемые информационные источники 24](#_Toc163052903)

[Приложения 25](#_Toc163052904)

# Введение

Цель настоящего исследования состоит в анализе потребительских предпочтений в различных регионах с целью выявления паттернов потребления продуктов питания. Для достижения этой цели была поставлена ряд задач:

Провести корреляционный анализ для изучения связи между различными продуктами питания.

Проанализировать результаты корреляционного анализа для выявления общих тенденций и взаимосвязей в потреблении продуктов.

Построить дендрограммы с использованием различных методов кластерного анализа для выделения групп регионов с схожими потребительскими предпочтениями.

Проанализировать полученные кластеры с целью выявления различий в уровне жизни, характеристиках производства и потребления продуктов.

В ходе исследования были использованы данные о потреблении различных продуктов питания в регионах, а именно: яйца, мясо, картофель, хлеб, молоко, овощи и фрукты. Эти показатели были выбраны как основные показатели потребления, способные отразить разнообразие диет и потребительских предпочтений в различных регионах.

На основе анализа данных были выдвинуты следующие рабочие гипотезы:

Потребление определенных продуктов питания часто происходит вместе, что будет отражено в положительных значениях коэффициентов корреляции между этими продуктами.

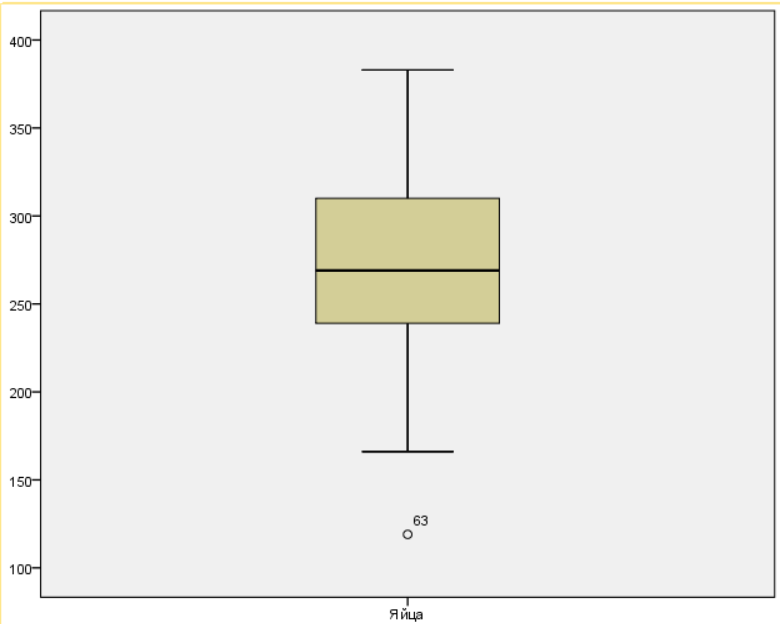
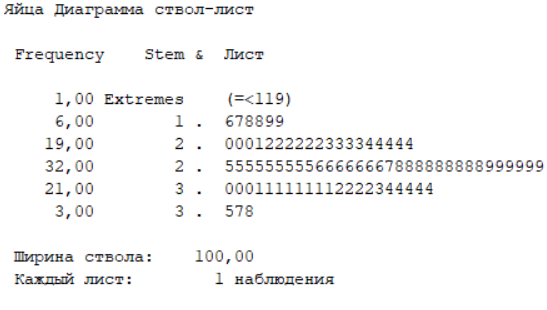
Существуют различные группы регионов с схожими потребительскими предпочтениями, что будет отражено в выделении кластеров при кластерном анализе.

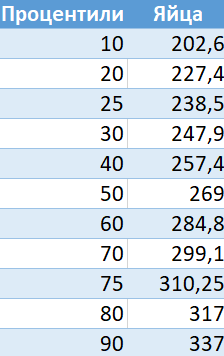
Потребление продуктов питания связано с уровнем жизни и характеристиками производства в регионах, что будет отражено в характеристиках выделенных кластеров.

# Предварительный анализ данных

Рассмотрим каждый показатель отдельно:

1. Яйца – Потребление яиц на душу населения в кг по различным регионам



Рис. 1 – Графики для яиц

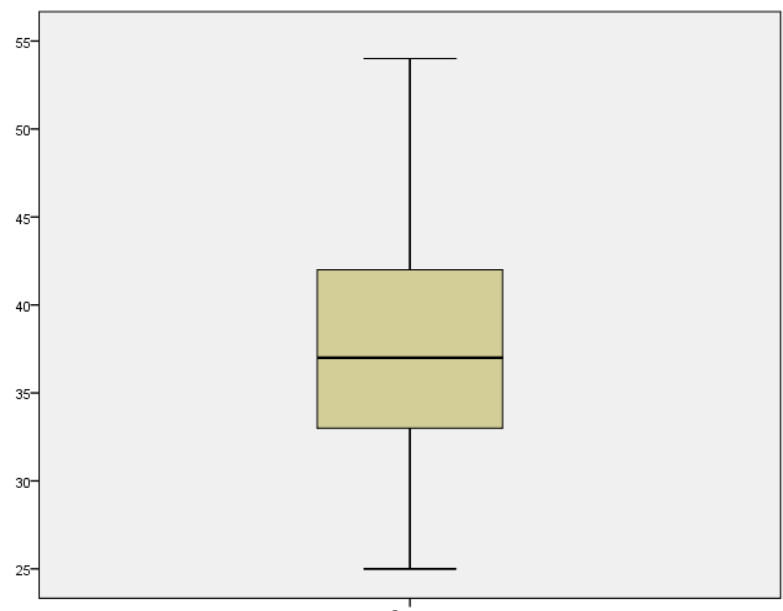
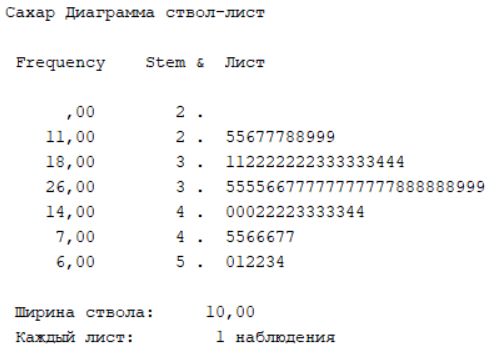
|  |  |
| --- | --- |
| Xср | 270,7683 |
| Me | 269 |
| Mo | 250 |
| R | 264 |
| СV | 0,186682 |
| S^2 | 2555,044 |
| S | 50,54745 |
| IQR | 71,75 |

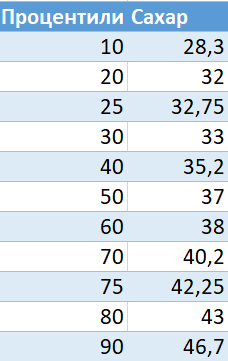
Выброс:

Республика Тыва 119

Из предоставленных данных следует, что средний уровень потребления яиц в кг на душу населения в рассматриваемых регионах составляет около 270,77 кг в год. Наиболее часто встречающееся значение потребления - 250 кг. Имеется небольшой разброс данных относительно среднего значения, что указывает на сходство в потреблении между регионами, но существуют и некоторые различия.

1. Сахар– Потребление сахара на душу населения в кг по различным регионам



Рис. 2 – Графики для сахара

|  |  |
| --- | --- |
| Xср | 37,5 |
| Me | 37 |
| Mo | 37 |
| R | 29 |
| СV | 0,182276 |
| S^2 | 46,72222 |
| S | 6,835366 |
| IQR | 9,5 |

Из предоставленных данных следует, что средний уровень потребления сахара на душу населения в рассматриваемых регионах составляет около 37,5 кг в год. Наиболее часто встречающееся значение потребления также составляет 37 кг. Разброс данных относительно среднего значения невелик, а стандартное отклонение равно примерно 6,84 кг, что указывает на низкую вариативность потребления между регионами. Однако, стоит отметить, что имеются различия в потреблении, хотя общие тенденции схожи.

1. Мясо– Потребление мяса на душу населения в кг по различным регионам

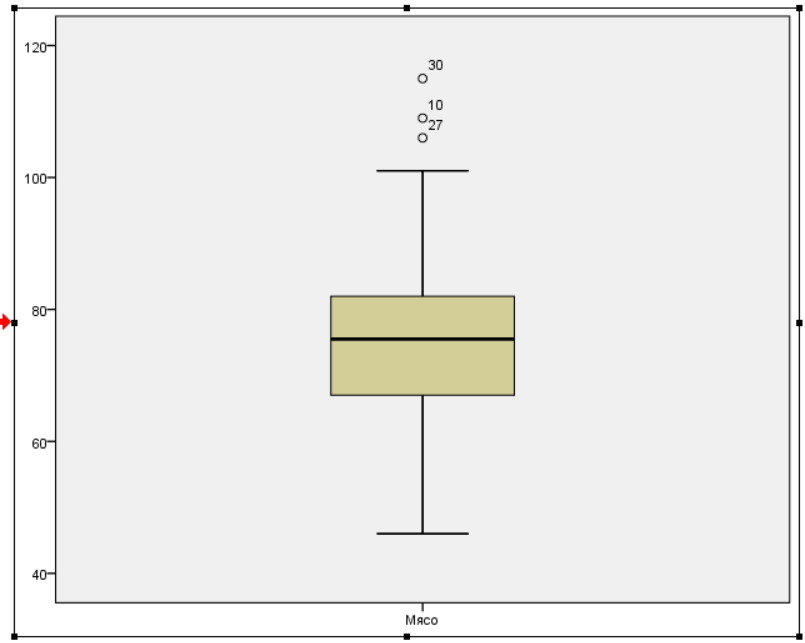
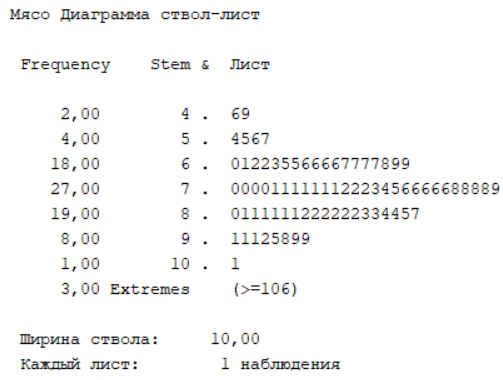
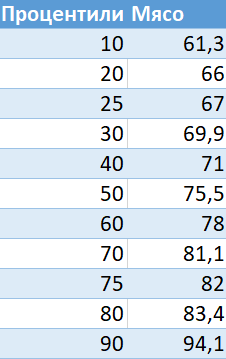


Рис. 3 – Графики для мяса

|  |  |
| --- | --- |
| Xср | 75,97561 |
| Me | 75,5 |
| Mo | 71 |
| R | 69 |
| СV | 0,170988 |
| S^2 | 168,7648 |
| S | 12,99095 |
| IQR | 15 |



Выброс:

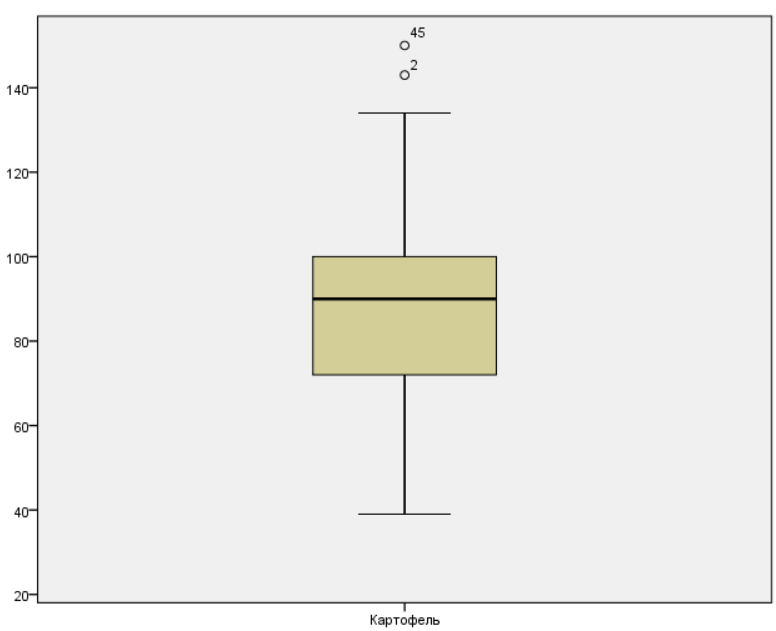
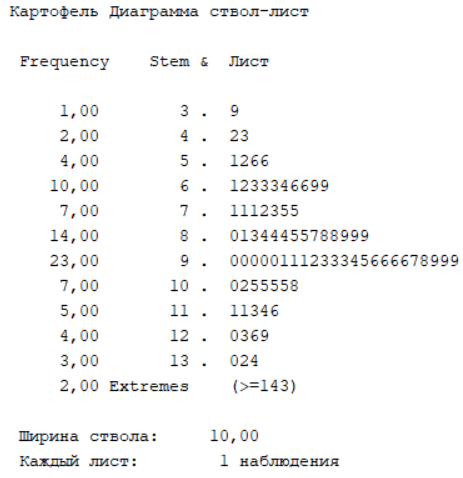
Республика Калмыкия 115

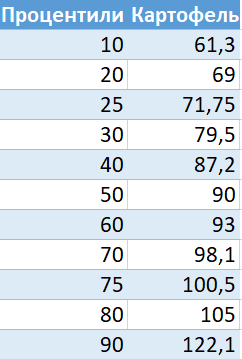
Разброс данных относительно среднего значения является умеренным, а стандартное отклонение составляет примерно 12,99 кг, что указывает на некоторую вариативность потребления мяса между регионами, но не очень большую.

Межквартильный размах составляет 15 кг, что означает, что 50% регионов имеют потребление мяса в диапазоне от приблизительно 60,5 кг до 75,5 кг.

Таким образом, общие тенденции показывают сходство в потреблении мяса между рассматриваемыми регионами, но имеются и некоторые различия.

1. Картофель– Потребление картофеля на душу населения в кг по различным регионам

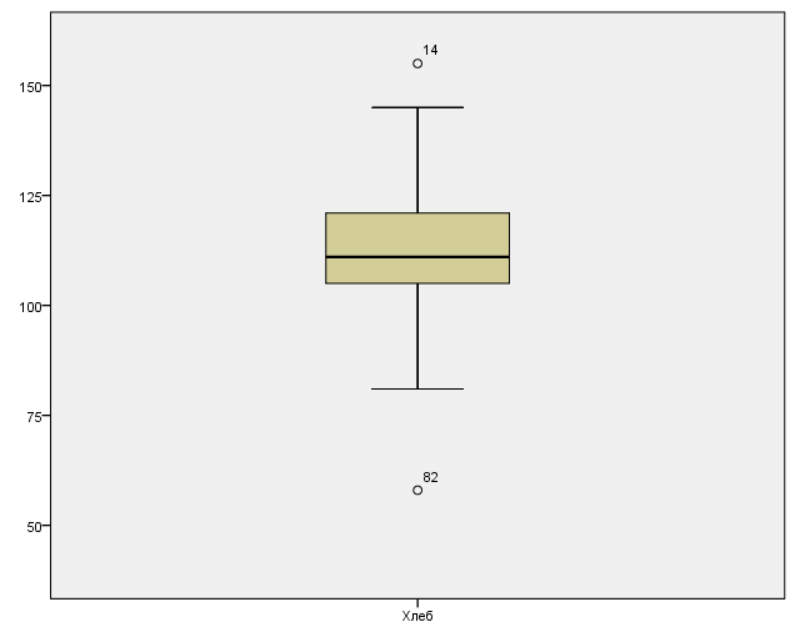
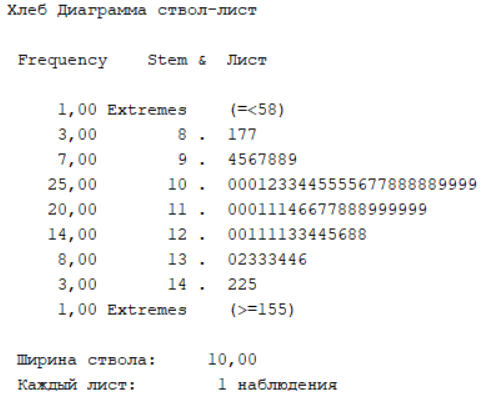


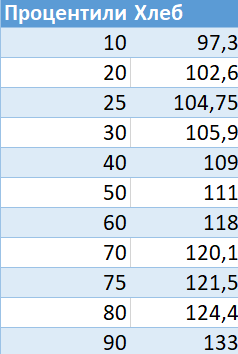
Рис. 4 – Графики для картофеля

|  |  |
| --- | --- |
| Xср | 89,3902439 |
| Me | 90 |
| Mo | 90 |
| R | 111 |
| СV | 0,256723207 |
| S^2 | 526,635953 |
| S | 22,94855013 |
| IQR | 28,75 |

Разброс данных относительно среднего значения является относительно высоким, а стандартное отклонение составляет примерно 22,95 кг, что указывает на значительную вариативность потребления картофеля между регионами. Межквартильный размах составляет 28,75 кг, что означает, что 50% регионов имеют потребление картофеля в диапазоне от приблизительно 74,25 кг до 103 кг. Следовательно, хотя средний уровень потребления картофеля в рассматриваемых регионах высок, наблюдается значительная вариативность этого показателя между регионами.

1. Хлеб– Потребление хлеба на душу населения в кг по различным регионам



Рис. 5 – Графики для хлеба

|  |  |
| --- | --- |
| Xср | 113,5122 |
| Me | 111 |
| Mo | 119 |
| R | 97 |
| СV | 0,133592 |
| S^2 | 229,9566 |
| S | 15,16432 |
| IQR | 16,75 |

Выброс:

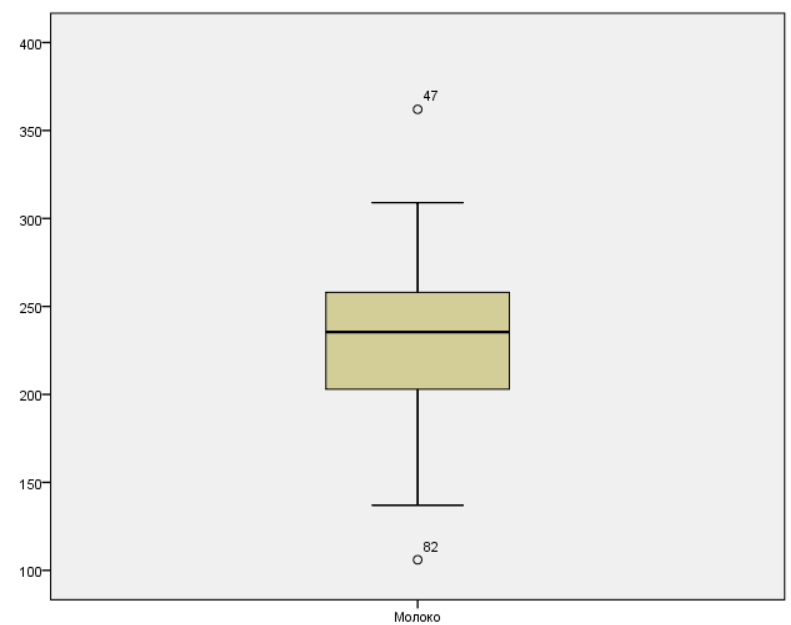
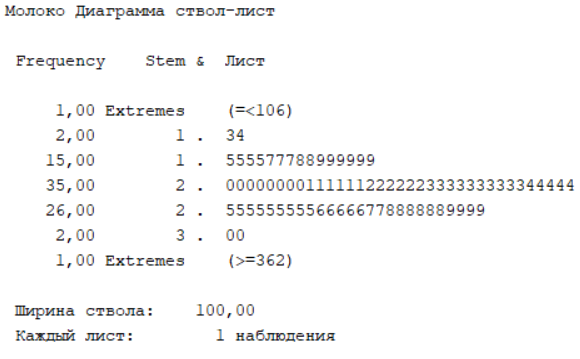
Чукотский автономный округ 58

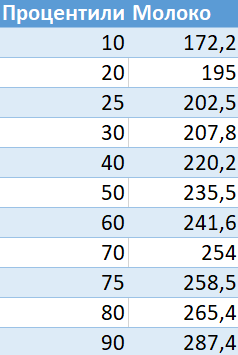
Разброс данных относительно среднего значения является умеренным, а стандартное отклонение составляет примерно 15,16 кг, что указывает на некоторую вариативность потребления хлеба между регионами, но не очень большую.

Межквартильный размах составляет 16,75 кг, что означает, что 50% регионов имеют потребление хлеба в диапазоне от приблизительно 98,25 кг до 115 кг.

Следовательно, общие тенденции показывают сходство в потреблении хлеба между рассматриваемыми регионами, хотя и имеются некоторые различия.

1. Молоко– Потребление молока на душу населения в кг по различным регионам



Рис. 6 – Графики для молока

|  |  |
| --- | --- |
| Xср | 230,4878 |
| Me | 235,5 |
| Mo | 254 |
| R | 256 |
| СV | 0,1940654 |
| S^2 | 2000,7468 |
| S | 44,729708 |
| IQR | 56 |

Разброс данных относительно среднего значения является умеренным, а стандартное отклонение составляет примерно 44,73 кг, что указывает на некоторую вариативность потребления молока между регионами, но не очень большую.

Межквартильный размах составляет 56 кг, что означает, что 50% регионов имеют потребление молока в диапазоне от приблизительно 199 кг до 255 кг.

Следовательно, общие тенденции показывают сходство в потреблении молока между рассматриваемыми регионами, хотя и имеются некоторые различия.

1. Овощи– Потребление овощей на душу населения в кг по различным регионам

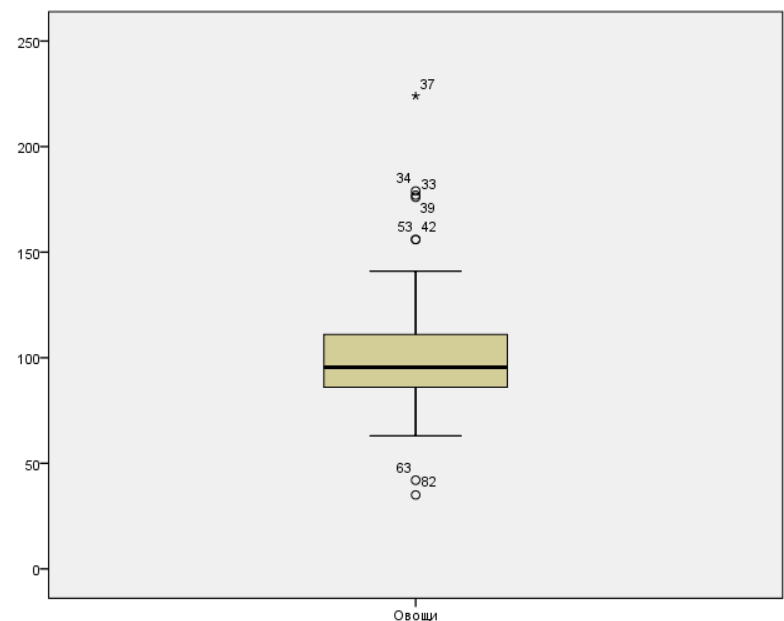
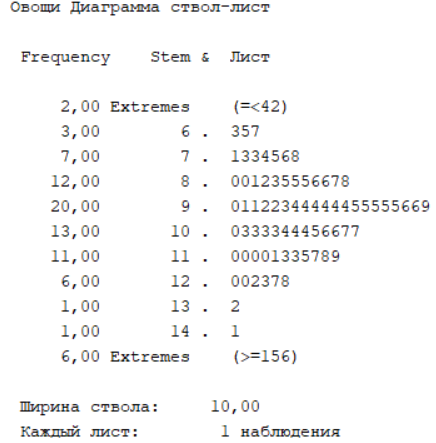
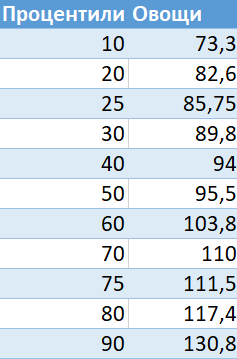


Рис. 7 – Графики для овощей

|  |  |
| --- | --- |
| Xср | 101,9634 |
| Me | 95,5 |
| Mo | 94 |
| R | 189 |
| СV | 0,283211 |
| S^2 | 833,8875 |
| S | 28,87711 |
| IQR | 25,75 |



Выброс:

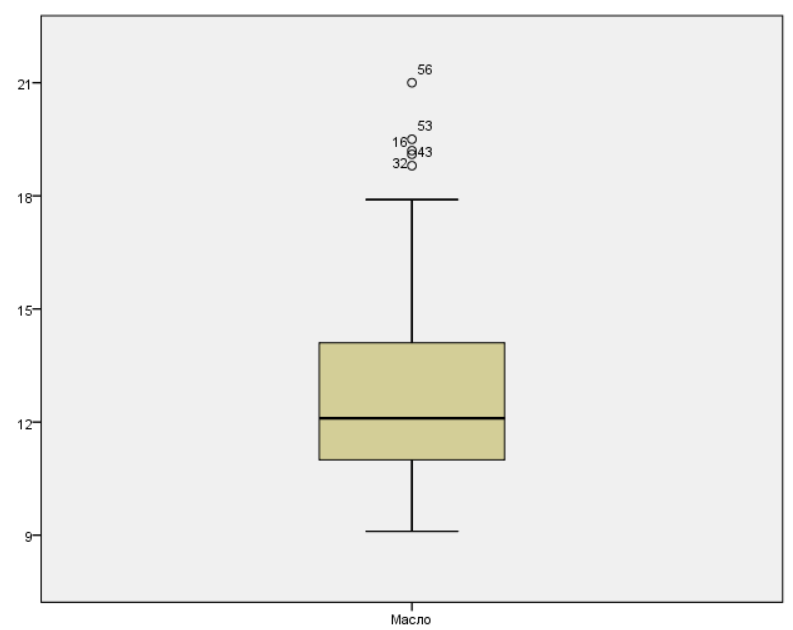
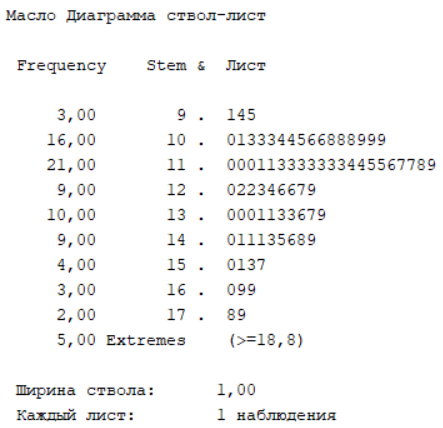
Республика Дагестан 224

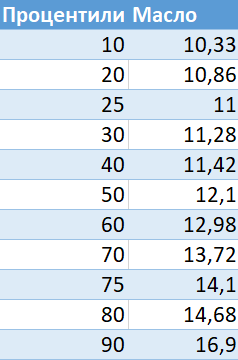
Разброс данных относительно среднего значения является высоким, а стандартное отклонение составляет примерно 28,88 кг, что указывает на значительную вариативность потребления овощей между регионами.

Межквартильный размах составляет 25,75 кг, что означает, что 50% регионов имеют потребление овощей в диапазоне от приблизительно 77,38 кг до 103,13 кг.

Следовательно, общие тенденции показывают, что потребление овощей в рассматриваемых регионах может быть довольно разнообразным, с высоким уровнем вариативности.

1. Масло– Потребление масла на душу населения в кг по различным регионам



Рис. 8 – Графики для масла

|  |  |
| --- | --- |
| Xср | 12,85244 |
| Me | 12,1 |
| Mo | 11,3 |
| R | 11,9 |
| СV | 0,20097 |
| S^2 | 6,671661 |
| S | 2,582956 |
| IQR | 3,1 |

Выброс:

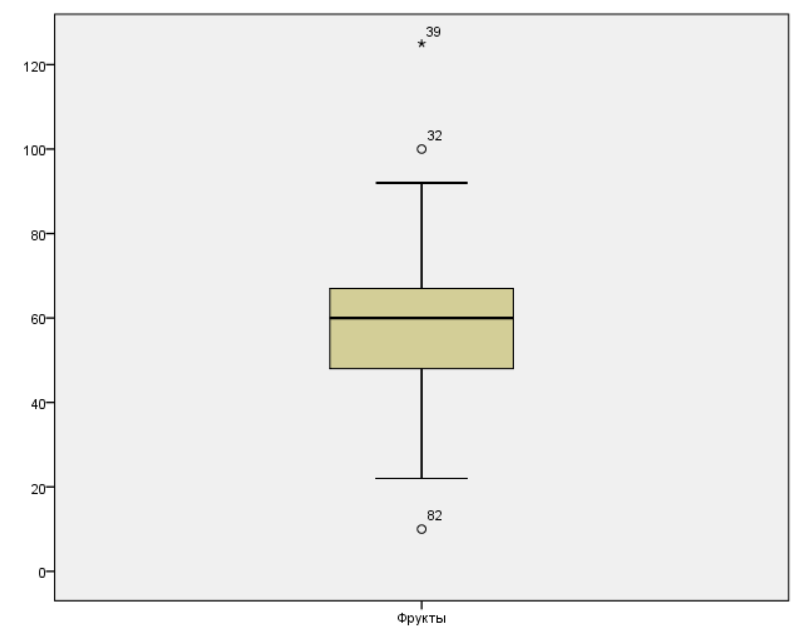
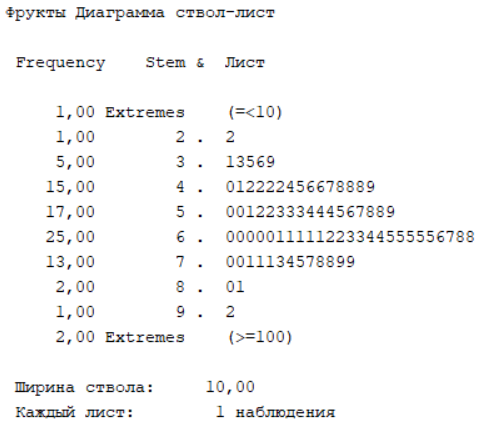
Саратовская область 21,0

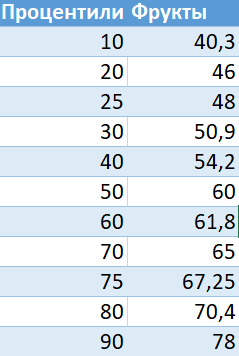
Разброс данных относительно среднего значения является умеренным, а стандартное отклонение составляет примерно 2,58 кг, что указывает на некоторую вариативность потребления масла между регионами, но не очень большую.

Межквартильный размах составляет 3,1 кг, что означает, что 50% регионов имеют потребление масла в диапазоне от приблизительно 10,1 кг до 13,2 кг.

Следовательно, общие тенденции показывают сходство в потреблении масла между рассматриваемыми регионами, хотя и имеются некоторые различия.

1. Фрукты– Потребление фруктов на душу населения в кг по различным регионам



Рис. 9 – Графики для фрутов

|  |  |
| --- | --- |
| Xср | 58,731707 |
| Me | 60 |
| Mo | 61 |
| R | 115 |
| СV | 0,2825255 |
| S^2 | 275,33454 |
| S | 16,593208 |
| IQR | 19,25 |

Выброс:

Кабардино-Балкарская Республика 125

Разброс данных относительно среднего значения является умеренным, а стандартное отклонение составляет примерно 16,59 кг, что указывает на некоторую вариативность потребления фруктов между регионами.

Межквартильный размах составляет 19,25 кг, что означает, что 50% регионов имеют потребление фруктов в диапазоне от приблизительно 50,88 кг до 70,13 кг.

Следовательно, общие тенденции показывают сходство в потреблении фруктов между рассматриваемыми регионами, хотя и имеются некоторые различия.

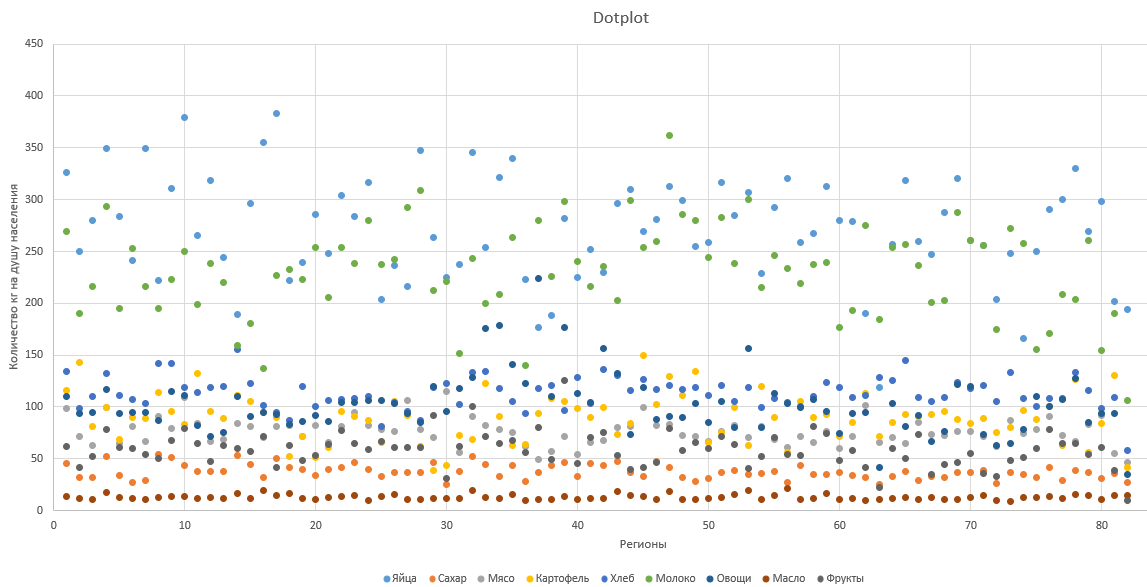


Рис. 9 – Точечное распределение для всех показателей

Из представленного анализа данных по потреблению различных продуктов на душу населения в кг в различных регионах можно сделать следующие общие выводы:

Потребление продуктов питания в рассматриваемых регионах имеет сходство в целом, но также присутствуют некоторые различия между ними.

По большинству продуктов (яйца, сахар, мясо, хлеб, молоко, масло, фрукты) средний уровень потребления на душу населения в рассматриваемых регионах находится в

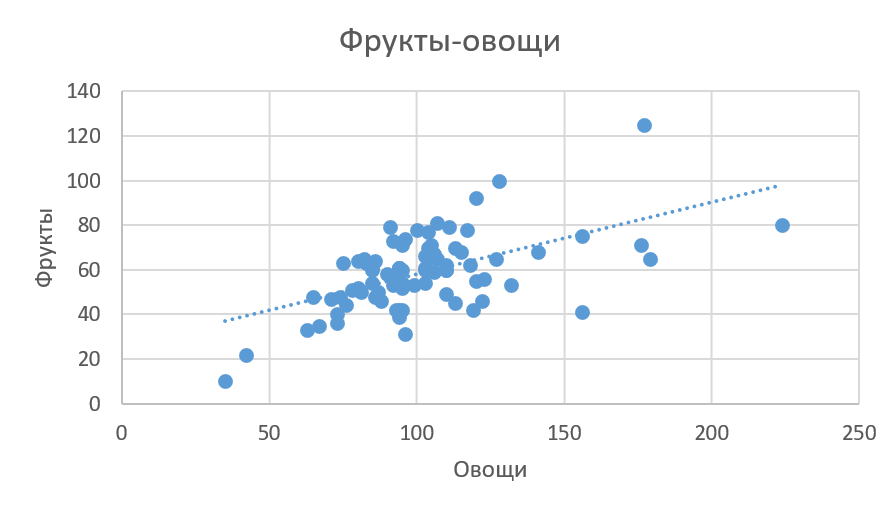
пределах нормы и не сильно различается между регионами.

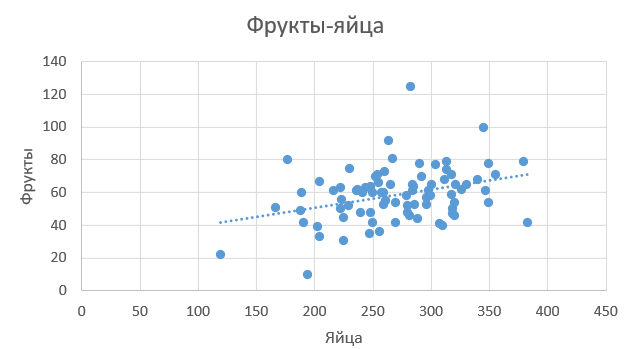
Вариативность потребления между регионами различна для каждого продукта. Например, потребление картофеля и овощей имеет высокий разброс данных, что может указывать на значительные культурные или экономические различия между регионами в использовании данных продуктов.

В некоторых регионах наблюдаются выбросы данных, указывающие на аномально высокие или низкие значения потребления определенных продуктов. Это может быть связано с особенностями местной экономики, культуры или пищевых привычек.

Таким образом, хотя существуют некоторые различия в потреблении продуктов между регионами, общие тенденции показывают сходство в пищевых привычках.

# Корреляционный анализ:

Построим поля корреляции для исследования связи между переменными до удаления выбросов:

Рис. 10 – Поле корреляции фруктов и овощей

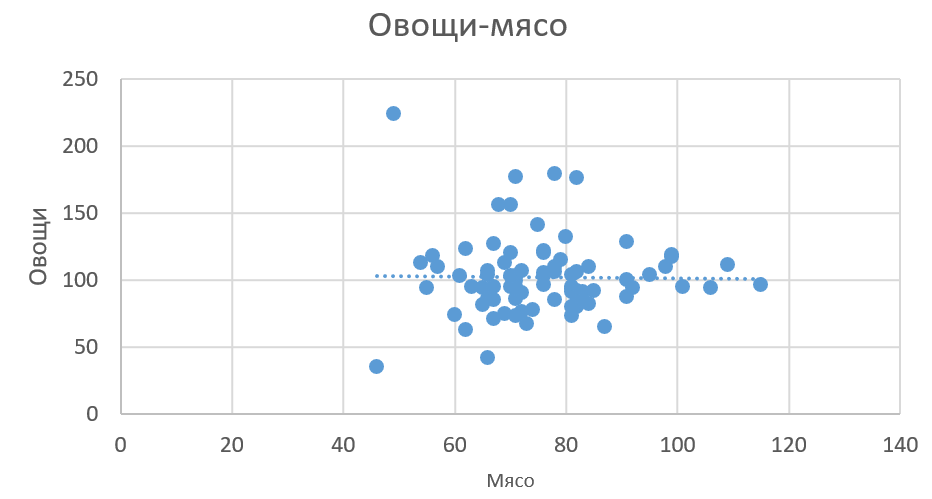
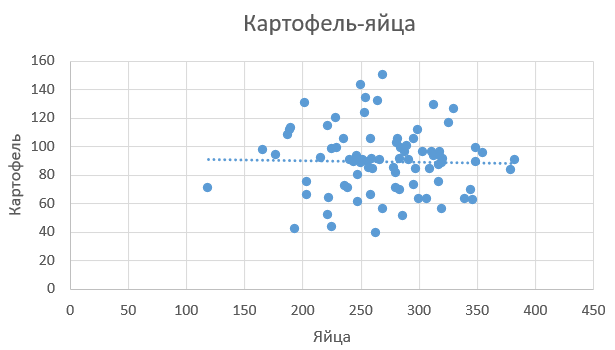
Рис. 11 – Поле корреляции фруктов и яиц

Рис. 12 – Поле корреляции мяса и овощей

Рис. 13 – Поле корреляции картофеля и яиц

(Остальные поля корреляции находятся в приложении к отчету.)

Опишем связи между переменными:

**Фрукты и Овощи:**

Между фруктами и овощами также наблюдается положительная средняя корреляция (корреляция = 0.562), что указывает на то, что потребление фруктов и овощей часто происходит вместе.

**Фрукты и яйца:**

Корреляция между фруктами и яйцами составляет 0.336, указывая на среднюю положительную связь между этими продуктами.

**Мясо и Овощи:**

Корреляция между мясом и овощами составляет -0.016, указывая на очень слабую обратную связь между этими продуктами. Обратная корреляция означает, что потребление мяса скорее связано с менее значительным потреблением овощей.

**Картофель и Яйца:**

Корреляция между картофелем и яйцами составляет -0.020, что указывает на очень слабую обратную связь между этими продуктами. Обратная корреляция означает, что потребление картофеля скорее связано с менее значительным потреблением яиц.

Остальные поля корреляции можно разбить на 2 группы:

**Корреляция 0.020 — 0.20**

Такая корреляция указывает на очень слабую связь между этими продуктами.

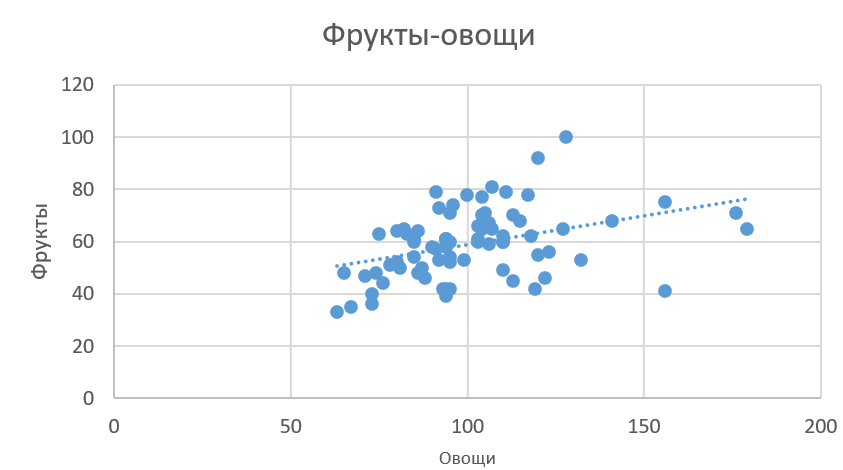
Значимость этой корреляции очень мала, что означает, что связь между потреблением этих продуктов не является статистически значимой.

**Корреляция 0.20 — 0.336**

Такая корреляция указывает на умеренную связь между этими продуктами.

Значимость этой корреляции значима.

**Поля корреляции после удаления аномальных выбросов**

Рис. 14 – Поле корреляции хлеба и картофеля

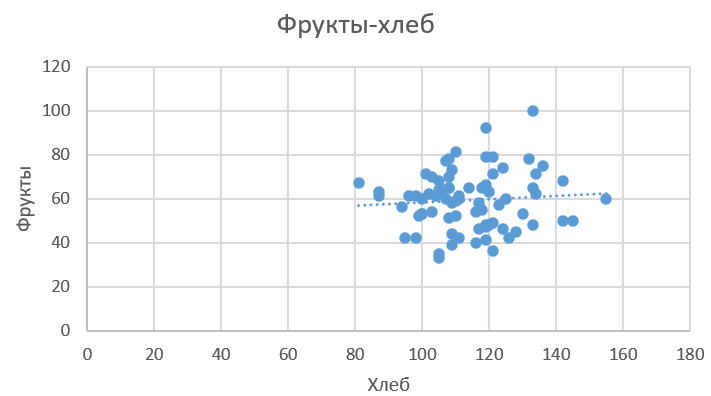
Рис. 15 – Поле корреляции фруктов и овощей

Рис. 16 – Поле корреляции фруктов и хлеба

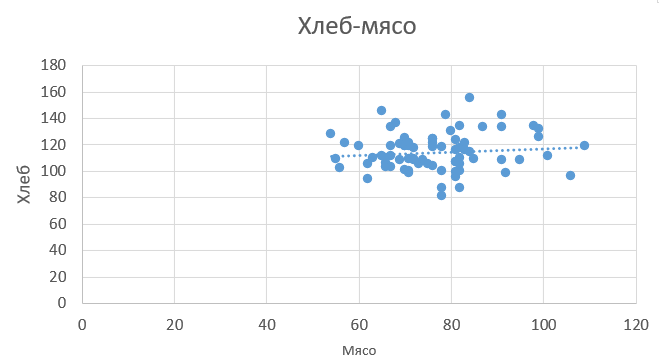


Рис. 17 – Поле корреляции хлеба и мяса

(Остальные поля корреляции находятся в приложении к отчету.)

Опишем связи между переменными:

**Фрукты и Овощи:**

Между фруктами и овощами также наблюдается положительная средняя корреляция (корреляция = 0.357), что указывает на то, что потребление фруктов и овощей часто происходит вместе.

Значимость этой корреляции также составляет 0.002, что подтверждает её статистическую значимость.

**Хлеб и картофель**

Корреляция между картофелем и другим продуктом составляет 0.289. Это указывает на среднюю положительную связь между картофелем и этим продуктом.

Значимость этой корреляции составляет 0.015, что означает, что связь статистически значима на уровне значимости 0.05.

**Мясо и хлеб**

Коэффициент корреляции составляет -0.015, что указывает на очень слабую обратную связь между мясом и хлебом.

Обратная корреляция означает, что потребление мяса скорее связано с менее значительным потреблением хлеба.

Значимость корреляции составляет 0.900. Это означает, что корреляция между мясом и хлебом не является статистически значимой на уровне значимости 0.05. Иными словами, нет убедительных доказательств того, что наблюдаемая связь является неслучайной.

**Хлеб и фрукты**

Корреляция между хлебом и фруктами составляет 0.011. Это указывает на крайне слабую положительную связь между этими двумя продуктами.

Значимость этой корреляции составляет 0.926, что говорит о том, что эта связь не является статистически значимой на уровне значимости 0.05.

Остальные поля корреляции можно разбить на 2 группы:

**Корреляция 0.015 — 0.15**

Такая корреляция указывает на очень слабую связь между этими продуктами.

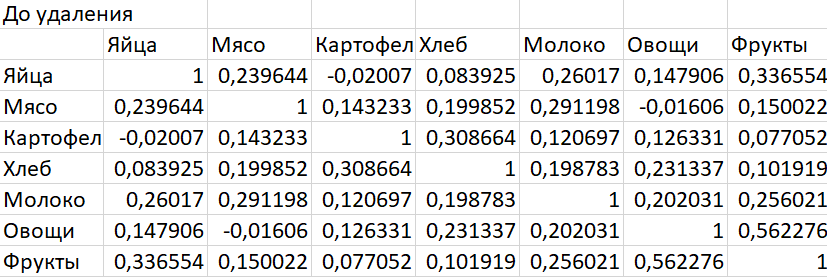
Значимость этой корреляции очень мала, что означает, что связь между потреблением этих продуктов не является статистически значимой.

**Корреляция 0.15 — 0.277**

Такая корреляция указывает на умеренную связь между этими продуктами.

Значимость этой корреляции мала, но все же является статистически значимой.

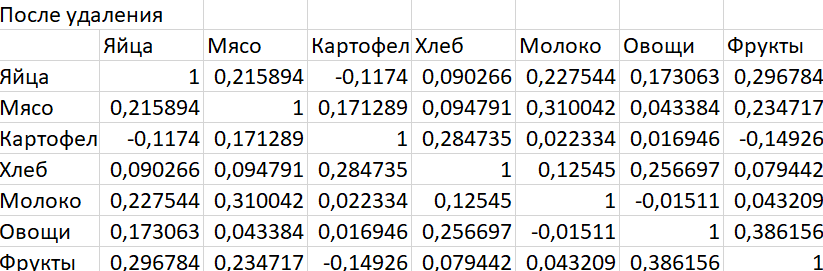
**Построим матрицы парных коэффициентов корреляции**

****

Коэффициенты корреляции между большинством переменных находятся в пределах от 0.1 до 0.3, что указывает на слабую связь между этими переменными.

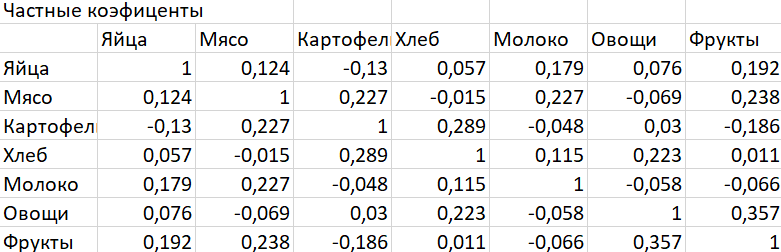
Наибольшие значения коэффициентов корреляции можно наблюдать между "Фруктами" и "Овощами" (0.562) и между "Фруктами" и "Яйцами" (0.337), что может указывать на некоторую положительную связь между этими продуктами.

Переменные "Овощи" и "Фрукты" имеют самый высокий коэффициент корреляции, что может указывать на наличие некоторой взаимосвязи в их потреблении.



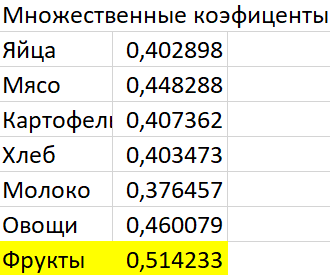
После удаления выбросов коэффициенты корреляции немного изменены, однако общий характер связей между переменными остается примерно таким же.

**Построим матрицу частных коэффициентов корреляции**



Произошло уменьшение связи между картофелем и другими продуктами, увеличение связи между мясом и молоком. После удаления влияния остальных переменных корреляция между фруктами и овощами остается высокой. Изменения в связи между яйцами и мясом(стала менее выраженной)

Данная матрица частных коэффициентов корреляции позволяет увидеть, как изменяются связи между переменными после удаления влияния других переменных. В данном случае видно, что у некоторых пар продуктов связь остается примерно на том же уровне или усиливается, в то время как у других связь становится слабее или даже меняет направление.

**Рассчитаем множественный коэффициент корреляции**

В целом, все продукты имеют некоторую связь между собой, но наиболее сильная связь наблюдается между потреблением овощей и фруктов, что указывает на их взаимозаменяемость или сходство в потребительских предпочтениях.

Итого, проведенное исследование демонстрирует, что существует определенная связь

между потреблением различных продуктов, но взаимосвязи могут быть сложными и зависеть от различных факторов, включая потребительские предпочтения, культурные особенности и диетические привычки.

# Кластерный анализ:

Построим дендрограммы, используя методы ближнего соседа, дальнего соседа, центроидный, связи, средней связи.

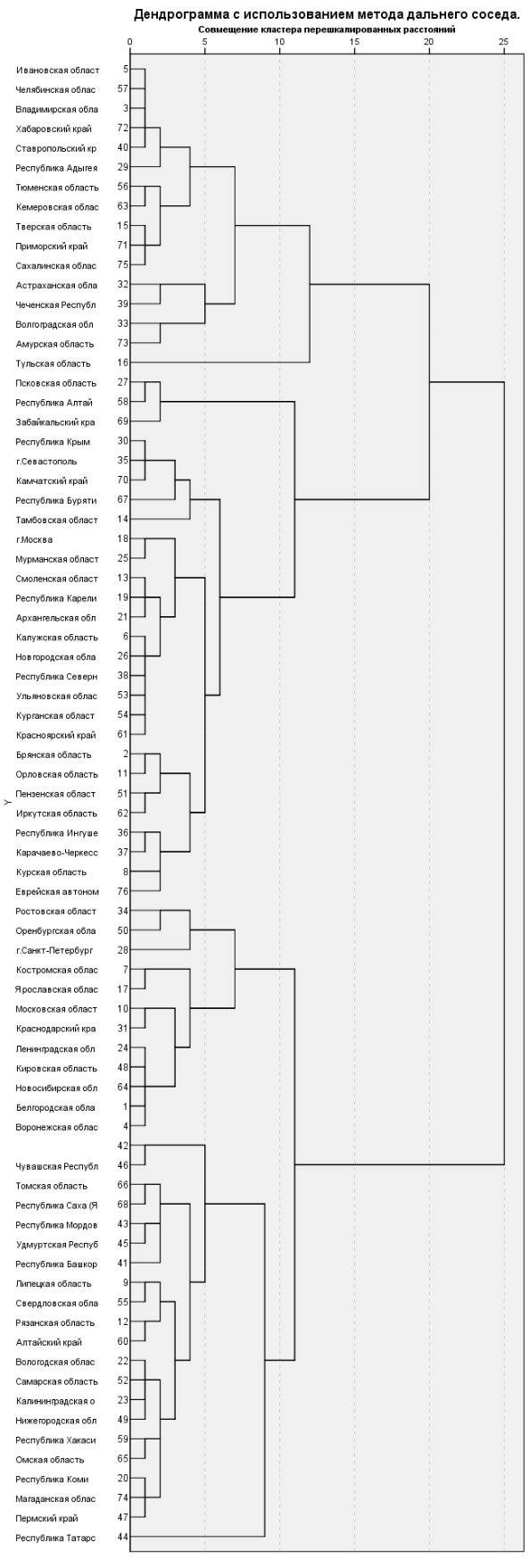
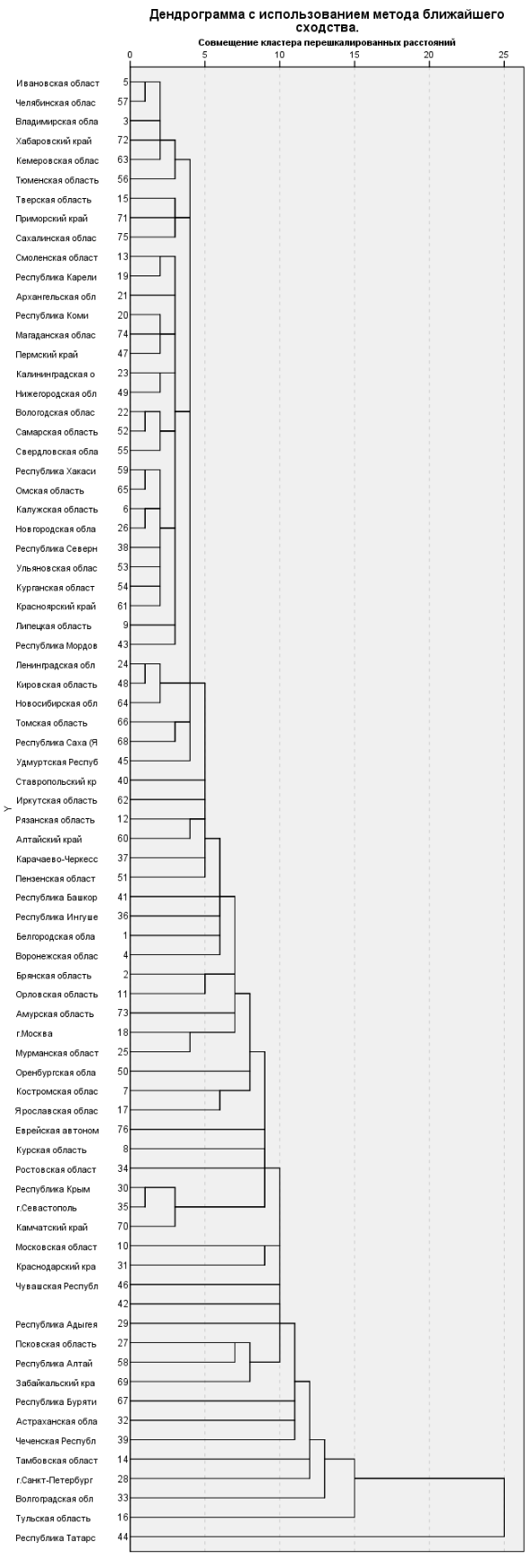


Рис. 18 – Дендрограммы

(Остальные дендрограммы находятся в приложенном файле)

Исходя из полученных дендрограмм, можно сделать вывод, что лучше всего делить на 3 группы.

Различные кластеры, выделенные в результате кластерного анализа, отражают разные уровни жизни в регионах. То есть один кластер включает регионы с более высоким уровнем жизни, где потребление определенных продуктов, таких как мясо, молоко и фрукты, более высокое, в то время как другие кластеры отражают регионы с более низким уровнем жизни и, следовательно, более скромным потреблением этих продуктов.

Также можно интерпретировать этот результат как деление на регионы, которые имеют схожие характеристики производства и потребления. То есть, некоторые регионы более ориентированы на сельское хозяйство и производство определенных продуктов в то время как другие регионы более индустриально развиты и иметь большее потребление продуктов, таких как мясо и молоко.

Таким образом, проведение кластерного анализа на основе данных о потреблении продуктов позволяет не только выявить общие тенденции в потреблении в различных регионах, но также исследовать различия в уровне жизни и характеристики производства и потребления в этих регионах.

На основании полученных результатов и дендрограммы, где оптимальное количество кластеров оценивается как 3, проведем кластеризацию методом k-средних как по нормализованным, так и по стандартизованным данным.

**По нормализированным данным:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Начальные центры кластеров** | | | |
|  | Кластер | | |
| 1 | 2 | 3 |
| Яйца | 313,00 | 188,00 | 355,00 |
| Мясо | 83,00 | 57,00 | 70,00 |
| Картофель | 129,00 | 108,00 | 95,00 |
| Хлеб | 121,00 | 121,00 | 101,00 |
| Молоко | 362,00 | 226,00 | 137,00 |
| Овощи | 91,00 | 110,00 | 95,00 |
| Фрукты | 79,00 | 49,00 | 71,00 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Конечные центры кластеров** | | | |
|  | Кластер | | |
| 1 | 2 | 3 |
| Яйца | 321,06 | 236,74 | 306,29 |
| Мясо | 82,35 | 74,84 | 75,14 |
| Картофель | 86,53 | 93,82 | 89,67 |
| Хлеб | 118,06 | 112,95 | 113,38 |
| Молоко | 279,41 | 224,03 | 205,43 |
| Овощи | 105,88 | 98,92 | 101,19 |
| Фрукты | 62,82 | 56,95 | 60,10 |

|  |  |  |
| --- | --- | --- |
| **Число наблюдений в каждом кластере** | | |
| Кластер | 1 | 17,000 |
| 2 | 38,000 |
| 3 | 21,000 |
| Валидные | | 76,000 |
| Пропущенные | | ,000 |

Теперь проанализируем характеристики каждого кластера на основе конечных центров:

Кластер 1:

Высокое потребление яиц и молока.

Среднее потребление мяса и овощей.

Низкое потребление фруктов и картофеля.

Самое высокое среди всех кластеров потребление молока.

Кластер 2:

Среднее потребление всех продуктов, ближе к среднему значению по всей выборке.

Близко к среднему значению по всем признакам, но немного ниже.

Кластер 3:

Высокое потребление яиц, мяса и молока.

Низкое потребление фруктов и картофеля.

Среднее потребление овощей.

|  |  |  |  |
| --- | --- | --- | --- |
| **Конечные центры кластеров** | | | |
|  | Кластер | | |
| 1 | 2 | 3 |
| Zscore(Яйца) | ,81707 | -,61871 | -,12515 |
| Zscore(Мясо) | ,53180 | -,31509 | -,11869 |
| Zscore(Картофель) | -,08788 | ,86143 | -,32437 |
| Zscore(Хлеб) | ,22087 | ,51221 | -,32260 |
| Zscore(Молоко) | ,30728 | -,72839 | ,16361 |
| Zscore(Овощи) | ,49257 | ,58466 | -,48245 |
| Zscore(Фрукты) | 1,14839 | -,21610 | -,45365 |

Итак, проведя анализ кластеров, мы видим, что каждый из них характеризуется определенным уровнем потребления различных продуктов. Кластер 1 выделяется высоким потреблением молока, кластер 2 представляет средний уровень потребления всех продуктов, а кластер 3 имеет высокий уровень потребления яиц, мяса и молока.

**По стандартизированным данным:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Начальные центры кластеров** | | | |
|  | Кластер | | |
| 1 | 2 | 3 |
| Zscore(Яйца) | 1,50387 | -1,83882 | ,75391 |
| Zscore(Мясо) | 1,23785 | ,63590 | ,37792 |
| Zscore(Картофель) | -1,00131 | ,90686 | -,31982 |
| Zscore(Хлеб) | 1,33645 | 2,90126 | ,12728 |
| Zscore(Молоко) | ,27406 | -1,68959 | 1,58316 |
| Zscore(Овощи) | 1,18170 | ,39082 | -1,23489 |
| Zscore(Фрукты) | 3,11249 | ,06614 | -1,45704 |

|  |  |  |
| --- | --- | --- |
| **Число наблюдений в каждом кластере** | | |
| Кластер | 1 | 19,000 |
| 2 | 17,000 |
| 3 | 40,000 |
| Валидные | | 76,000 |
| Пропущенные | | ,000 |

Кластер 1:

Более высокое потребление яиц, мяса, молока, хлеба и фруктов.

Низкое потребление картофеля и овощей.

Кластер 2:

Низкое потребление всех продуктов.

Близко к среднему значению по всем признакам, но немного ниже.

Кластер 3:

Низкое потребление яиц, мяса, картофеля, хлеба, овощей и фруктов.

Более высокое потребление молока.

Анализируя характеристики каждого кластера, мы видим, что кластеры, сформированные по стандартизированным данным, имеют более четко выраженные различия в потреблении продуктов, по сравнению с кластерами, сформированными по нормализованным данным.

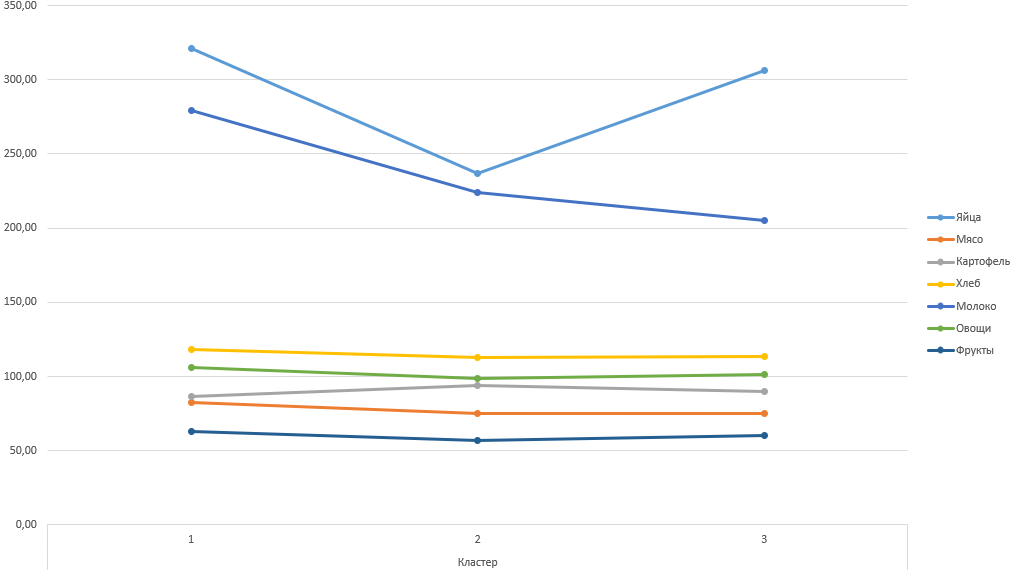
При сравнении результатов кластеризации по нормализованным и стандартизованным данным, видно, что они значительно различаются.

В данном случае, кластеризация по стандартизованным данным кажется более интерпретируемой и осмысленной. Кластеры, сформированные по стандартизованным данным, выделяют более явные паттерны потребления продуктов: высокое, среднее и низкое потребление. Это может быть обусловлено тем, что стандартизация данных учитывает не только абсолютные значения, но и отклонения от среднего, что может быть важным при анализе таких данных.

Интерпретация результатов кластерного анализа позволяет сделать вывод о различиях в уровне жизни и характеристиках производства и потребления в регионах, представленных в каждом кластере. Например, регионы, входящие в кластер с высоким потреблением молока и яиц, могут иметь более высокий уровень жизни и быть более ориентированными на сельское хозяйство.

Сравнение результатов кластеризации по нормализованным и стандартизованным данным показывает, что стандартизация данных может привести к более четко выраженным различиям между кластерами и более интерпретируемым результатам.

Таким образом, кластерный анализ позволяет не только выделить группы схожих регионов на основе их потребительских предпочтений, но и провести интерпретацию полученных кластеров с точки зрения уровня жизни, характеристик производства и потребления, что может быть полезным для принятия решений в области планирования и развития региональных рынков продуктов питания.

Рис. 19 – График средних значений кластеров нормализованные данные

Кластер 1: "Высокое потребление молока и яиц"

Яйца и молоко потребляются в больших количествах по сравнению с другими продуктами.

Среднее потребление мяса и овощей.

Низкое потребление фруктов и картофеля.

Кластер 2: "Сбалансированное потребление"

Среднее потребление всех продуктов, ближе к среднему значению по всей выборке.

Кластер 3: "Высокое потребление яиц, мяса и молока"

Высокое потребление яиц, мяса и молока.

Низкое потребление фруктов и картофеля.

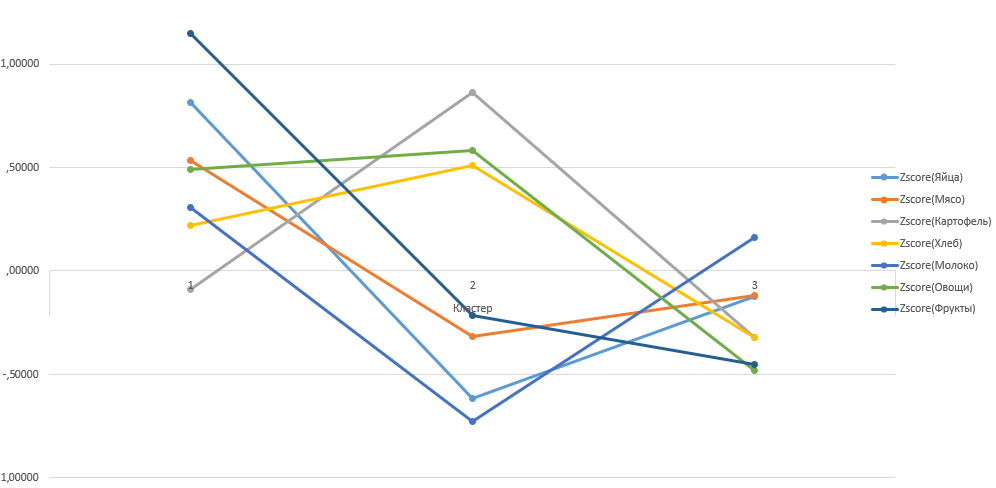
Среднее потребление овощей.

Рис. 20 – График средних значений кластеров стандартизованные данные

Кластер 1: "Большой разброс в потреблении"

Большой разброс в потреблении продуктов, но относительно высокое потребление яиц, мяса, молока, хлеба и фруктов.

Низкое потребление картофеля и овощей.

Кластер 2: "Низкое потребление всех продуктов, за исключением картофеля"

Низкое потребление всех продуктов, кроме картофеля. Его потребление сильно больше по сравнению с другими продуктами

Близко к среднему значению по всем признакам, но немного ниже.

Кластер 3: "Высокое потребление молока"

Низкое потребление яиц, мяса, картофеля, хлеба, овощей и фруктов.

Более высокое потребление молока по сравнению с другими продуктами.

Интерпретация результатов:

По результатам кластерного анализа, мы выделили три основных типа потребления продуктов:

Кластер с высоким потреблением молока и яиц: Этот кластер может отражать регионы или группы людей, где предпочтительнее потребление молочных и яичных продуктов, возможно, из-за их доступности или культурных предпочтений.

Кластер с сбалансированным потреблением: Этот кластер представляет собой средний уровень потребления всех продуктов, что может указывать на типичное потребление в среднем регионе или среди среднестатистической группы людей.

Кластер с высоким потреблением яиц, мяса и молока: Этот кластер может отражать регионы или группы людей с высоким уровнем потребления белковых продуктов, что может быть связано с особенностями диеты или доступности этих продуктов.

Выводы:

Кластерный анализ позволил выявить различные паттерны потребления продуктов в регионах или среди групп людей. Результаты подтверждают, что потребление продуктов может сильно варьироваться в зависимости от различных факторов, таких как культурные особенности, доступность продуктов, диетические привычки и другие. Эти результаты могут быть полезны для разработки целевых программ по питанию, сельскому хозяйству и здравоохранению, а также для лучшего понимания потребительских предпочтений и потребностей.

# Общие выводы по работе

Целью настоящего исследования было анализировать потребительские предпочтения в различных регионах с целью выявления паттернов потребления продуктов питания. Для достижения этой цели был выполнен ряд задач, включая проведение предварительного, корреляционного анализа и кластерного анализа на данных о потреблении различных продуктов.

Предварительный анализ данных позволил оценить основные характеристики и распределения переменных, а также выявить потенциальные аномалии или пропущенные значения. Этот этап анализа позволяет глубже понять данные и подготовить их для более детального исследования. В данной работе предварительный анализ помог удостовериться в качестве данных и их пригодности для дальнейшего корреляционного и кластерного анализа. Исходя из предварительного анализа было сделано предположение о том, что данные достаточно чистые и подготовлены для дальнейшего анализа. Анализ распределений переменных позволил определить их характеристики, а матрица корреляции выявила потенциальные взаимосвязи между продуктами питания. Эти результаты послужили основой для последующего корреляционного и кластерного анализа, позволяя более глубоко исследовать потребительские предпочтения в различных регионах.

В ходе корреляционного анализа были выявлены сильные и слабые связи между различными продуктами питания. Наиболее сильная положительная связь была обнаружена между потреблением фруктов и овощей, что подтверждает их взаимозаменяемость в диете. Это соответствует рабочей гипотезе о том, что потребление определенных продуктов часто происходит вместе.

Кластерный анализ позволил выделить три основных типа потребления продуктов: с высоким, сбалансированным и средним уровнем потребления. Различия между кластерами подтверждают гипотезу о существовании различных групп регионов с схожими потребительскими предпочтениями.

Анализ выделенных кластеров показал, что потребление продуктов питания связано с уровнем жизни и характеристиками производства в регионах. Это подтверждает гипотезу о том, что потребление продуктов питания может быть индикатором социоэкономических и производственных характеристик регионов.

Таким образом, результаты исследования подтверждают наши предположения о взаимосвязи потребления продуктов питания с социоэкономическими и производственными факторами в регионах. Эти результаты могут быть полезны для разработки целенаправленных стратегий по питанию и сельскому хозяйству, а также для оценки уровня жизни и потребительских предпочтений в различных регионах.

# Используемые информационные источники

Сайт системы ЕМИСС – [www.fedstat.ru](http://www.fedstat.ru)

Федеральная служба государственной статистики - <https://rosstat.gov.ru>

# Приложения

Приложением к модульной работе 1 является файл Статанализ.rar

В него входит:

Исходные файлы с данными:

Потребление фруктов.xls

Потребление картофеля.xls

Потребление масла.xls

Потребление молока.xls

Потребление мяса.xls

Потребление сахара.xls

Потребление хлебных продуктов.xls

Потребление сахара.xls

Потребление овощей.xls

Потребление яиц.xls

Вывод кластеризация k.spv-файл с результатами работы с кластеризацией

Вывод частная корреляция.spv-файл с результатами работы с корреляцией

Вывод кластеры.spv-файл с дендрограммами

Вывод boxplot+stemplot+1.5,3 IQR.spv-файл с графиками boxplot и stemplot

СР1.sav – файл с данными и переменными с выбросами

СР1\_без\_выбросов.sav - файл с данными и переменными без выбросов

СР1.xls – Excel файл с различными расчетами