Пермский филиал федерального государственного автономного образовательного учреждения высшего образования   
«Национальный исследовательский университет   
«Высшая школа экономики»

*Факультет экономики, менеджмента и бизнес-информатики*

Можегова Анна Сергеевна

**Модульное задание №1 по дисциплине**

**«Классификация статистических данных»**

студента образовательной программы «Программная инженерия»

по направлению подготовки *09.03.04 Программная инженерия*

**Руководитель:**

Тихонова Арина  
 Михайловна

Пермь, 2023 год

Оглавление

[Введение 3](#_Toc131275660)

[Глава 1 Предварительный анализ данных 5](#_Toc131275661)

[1.1 Экономическая интерпретация 5](#_Toc131275662)

[1.2 Характеристики по шкалам 5](#_Toc131275663)

[1.3 Графическое представление исходных данных 5](#_Toc131275664)

[1.3.1 Точечное распределение 6](#_Toc131275665)

[1.3.2 Листовая диаграмма 10](#_Toc131275666)

[1.3.3 Диаграмма выбросов 14](#_Toc131275667)

[1.4 Характеристики положения и разброса 20](#_Toc131275668)

[1.4.1 Средняя, медиана и мода 20](#_Toc131275669)

[1.4.2 Размах вариации, коэффициент вариации, дисперсия и стандартное отклонение 22](#_Toc131275670)

[1.5 Ранговые характеристики 24](#_Toc131275671)

[1.5.1 Квартили и децили 25](#_Toc131275672)

[1.5.2 Межквартильный размах, правила для определения выбросов 26](#_Toc131275673)

[1.5.2.1 1.5 IQR 26](#_Toc131275674)

[1.5.2.2 3 сигмы 28](#_Toc131275675)

[1.6 Z-преобразование 29](#_Toc131275676)

[1.7 Выводы 31](#_Toc131275677)

[Глава 2 Корреляционный анализ 32](#_Toc131275678)

[2.1 Построение полей корреляции 32](#_Toc131275680)

[2.2 Описание связи между переменными 35](#_Toc131275681)

[2.3 Матрица парных коэффициентов корреляции 37](#_Toc131275682)

[2.4 Построение полей корреляции после удаления выбросов 39](#_Toc131275683)

[2.5 Матрица парных коэффициентов корреляции после удаления выбросов 42](#_Toc131275684)

[2.6 Матрица частных коэффициентов корреляции 43](#_Toc131275685)

[2.7 Сравнение частных и парных коэффициентов корреляции 44](#_Toc131275686)

[2.8 Множественный коэффициент корреляции 44](#_Toc131275687)

[2.9 Вывод 45](#_Toc131275688)

[Глава 3 Кластерный анализ 46](#_Toc131275689)

[3.1 Разбиение на кластеры различными способами 46](#_Toc131275691)

[3.2 Выбор предпочтительно числа кластеров 60](#_Toc131275692)

[Вывод 61](#_Toc131275693)

[Приложение 62](#_Toc131275694)

Введение

В представленной работе выполняется корреляционный и кластерный анализ, а также предварительный анализ данных. В качестве объекта исследования были взяты данные об оценке туристического потока, оборот общественного питания на душу, валовая добавленная стоимость туристической индустрии, среднемесячная заработная плата по полному кругу организация, численность размещенных лиц в коллективных средствах размещения, валовой региональный продукт на душу населения и число турфирм за 2022 год.

Обязательно нужно проводить предварительный анализ поскольку именно на это этапе определяется однородность данных, их распределение и выбросы. В дальнейшем будет приниматься решение об удалении выбросов.

Необходимость проведения корреляционного анализа обосновывается тем, что этот статистический метод позволяет с использованием коэффициентов корреляции определить, существует ли зависимость между переменными и насколько она сильна. Как известно туристический поток зависит от уровня туристской активности, определяемой состоянием спроса и предложения туристских продуктов, а также политикой государства в этой сфере, но существует множество факторов, которые могли бы влиять на этот показатель. А именно, есть предположение, что оценка туристического потока имеет прямую зависимость от оборота общественного питания, среднемесячной заработной платы по полному кругу организация, численности размещенных лиц в коллективных средствах размещения, валовой региональный продукт на душу населения и числа турфирм, а обратную связь с валовой добавленной стоимостью туристической индустрии экономики.

Кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу, были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга. В данном случае кластерный анализ будет полезен для выявления аномалий, т.е. после кластеризации могут появиться отдельные данные, которые не относятся ни к одному из кластеров. Их нужно изучить, чтобы понять, ошибка это или какой-то интересный феномен, а также будет неплохо сжать имеющиеся данные, поделив их на кластеры, усреднив и оставив по одному объекту на каждый кластер. Это позволит в дальнейшем при анализе использовать меньше мощности.

Целью работы является проведение всех перечисленных анализов имеющихся данных.

Задачами, которые необходимо выполнить для достижения поставленной цели, являются:

* изучение выбранных данных;
* выявление и удаление показателей, относящийся к выбросам;
* описание распределения данных;
* рассмотрение графиков (графическое представление);
* расчет и анализ основных характеристик (характеристики положения, характеристики разброса и ранговые характеристики);
* построение полей корреляции до и после удаления выбросов;
* описание связи между переменными;
* построение матриц парных и частных коэффициентов корреляции;
* сравнение парных и частных коэффициентов;
* расчет множественного коэффициента корреляции;
* провести разбиение на кластеры.

# **Глава 1 Предварительный анализ данных**

В данной главе производится описание показателей, рассматриваются различные графические представления данных, рассчитываются некоторые характеристики совокупности, все это проделывается для дальнейшего выявления выбросов и каких-либо интересных факторов.

* 1. Экономическая интерпретация

Для своего исследования в качестве исходных данных я брала данные по регионам, а именно оценку туристического потока, оборот общественного питания на душу, валовая добавленная стоимость туристической индустрии, среднемесячная заработная плата по полному кругу организация, численность размещенных лиц в коллективных средствах размещения, валовой региональный продукт на душу населения и число турфирм(см. Вложение 1).

* 1. Характеристики по шкалам

Названия регионов является номинальной шкалой, ведь это данные, с которыми мы не может делать каких-либо арифметических операций, а вот все остальные — это интервальные шкалы, так как есть возможность находить разницу между двумя величинами, таким образом, мы, например, можем рассмотреть разность турфирм 2х различных регионов.

* 1. Графическое представление исходных данных

Рассмотрим каждый показатель по отдельности и изобразим данные на графиках, в качестве представления возьмём точечное распределение (Dotplot), листовую диаграмму (Stemplot), ящичковую диаграмму (Boxplot).

На основе точечного распределения данных можно предположить какие значения являются выбросами, диаграмма стебель-листья удобна для краткого обзора распределения, по ней также можно предположить о выбросах и моде, диаграмма ящик с усами однозначно показывает распределение и выбросы.

По оси X будем располагать значения от 1 до 85, так были пронумерованы все регионы, а по оси Y откладываем показатели тех или иных значений, никакой предварительной сортировки данных не производим.

* + 1. Точечное распределение

Точечный график для оценки туристического потока изображен на рисунке 1.3.1.1, по нему видно, что основная часть значений из группы данных находится в интервале от 0 до 250000, выбросами можно посчитать показатели регионов, которые выходят за этот интервал, по моему мнению, очевидным выбросом является регион, чья оценка туристического потока больше 1250000.

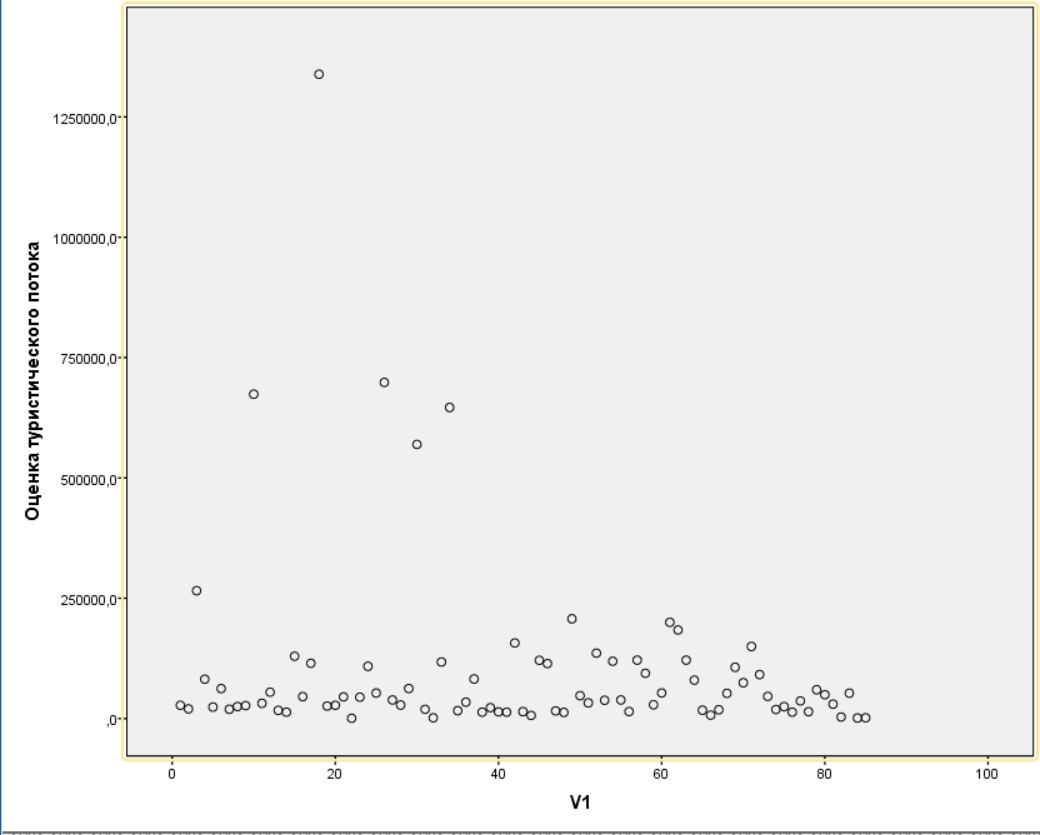
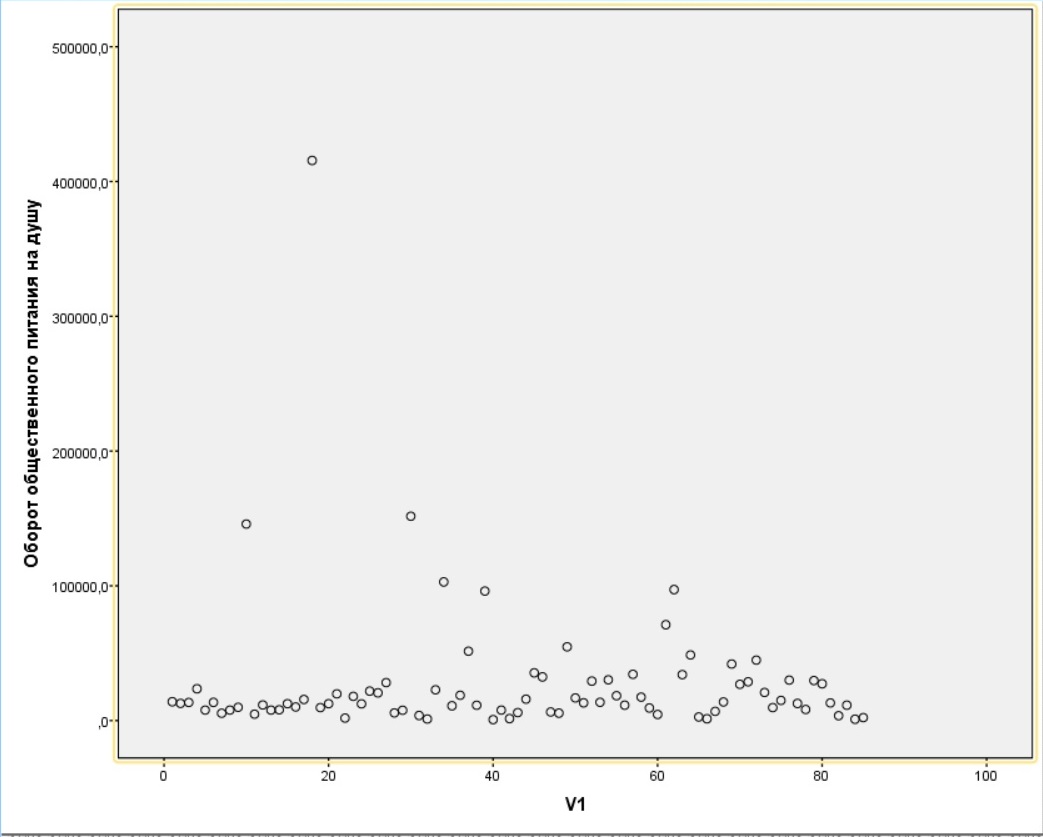


Рисунок 1.3.1.1 – Точечный график оценки туристического потока

Далее рассмотрим оборот общественного питания, мы получили график похожий на предыдущий, поскольку видно, что в данном примере очевидным выбросом оказался тот же регион, что и в показателях оценки туристического потока. Основная масса регионов имеет оборот в интервалах от 0 до 100000. (рис.1.3.1.2)

 Рисунок 1.3.1.2 – Точечный график оборота общественного питания

На рисунке 1.3.1.3 показан точечный график для значений валовой добавленной стоимости туристической индустрии экономики. Пока что наблюдается стабильная картина, где есть один и тот же регион, который скорее всего окажется выбросом и еще 3 региона, которые предположительно могут оказаться в списке выбросов. Основная масса показателей расположилась в интервале от 0 до 100000.

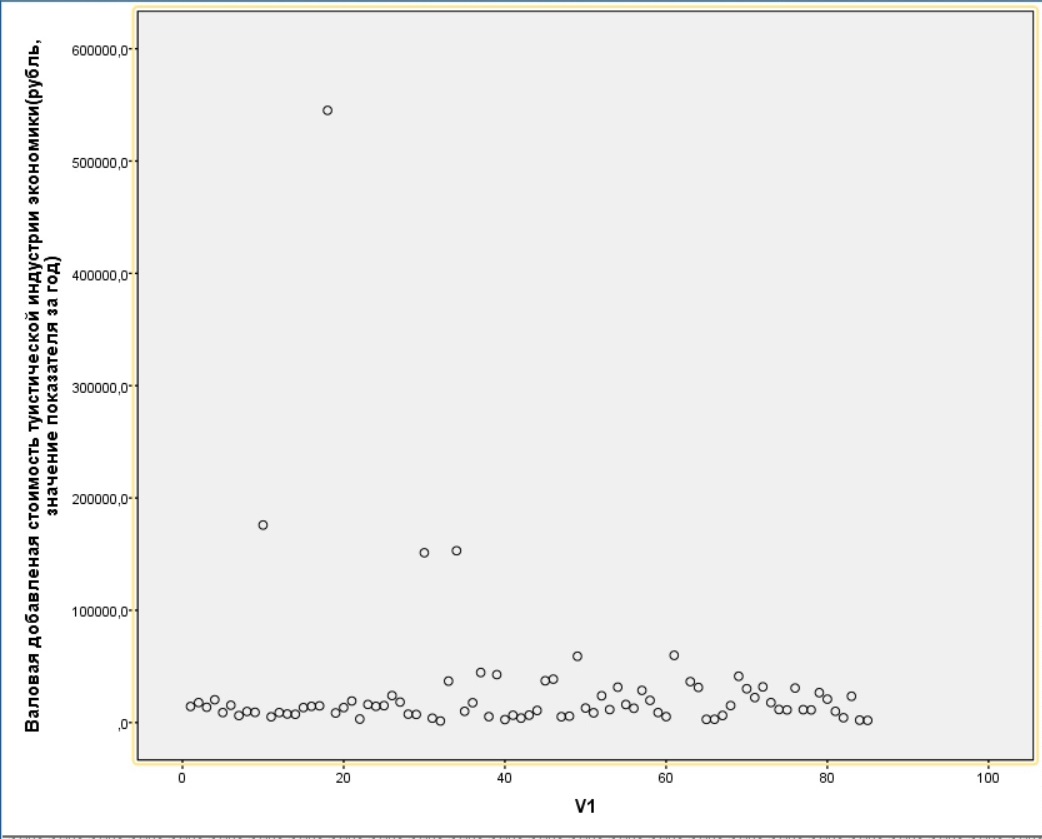
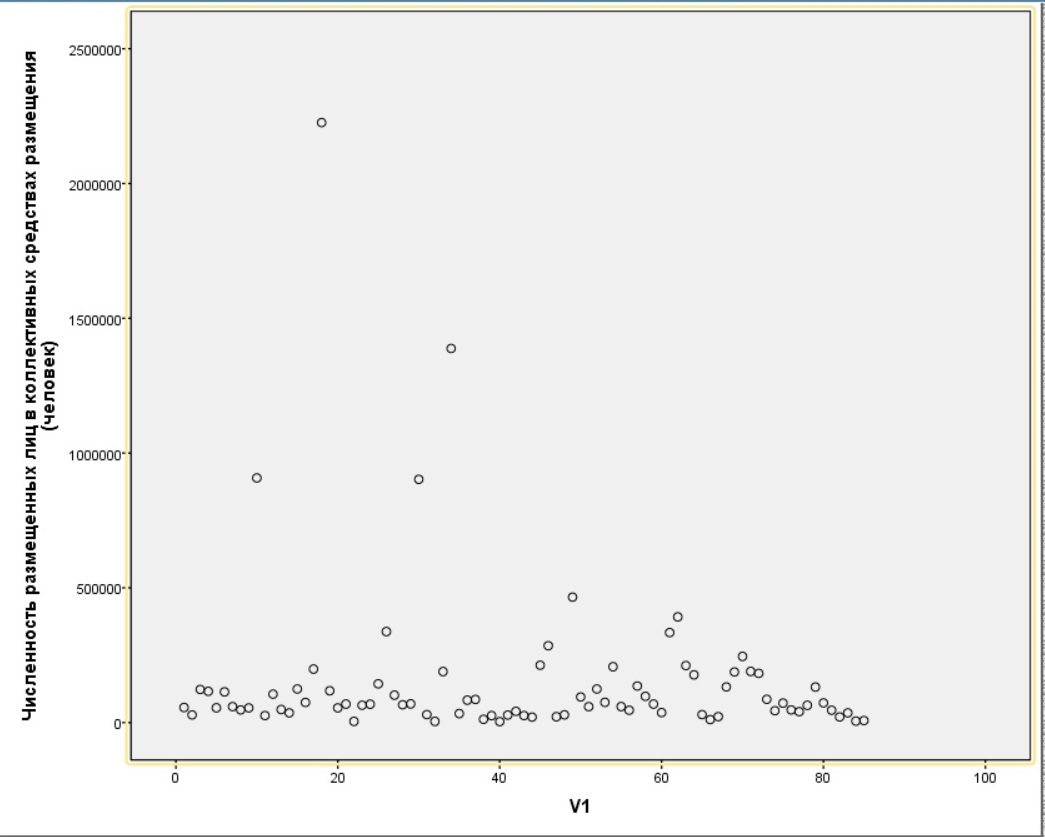
 Рисунок 1.3.1.3 – Точечный график валовой добавленной стоимости туристической индустрии экономики

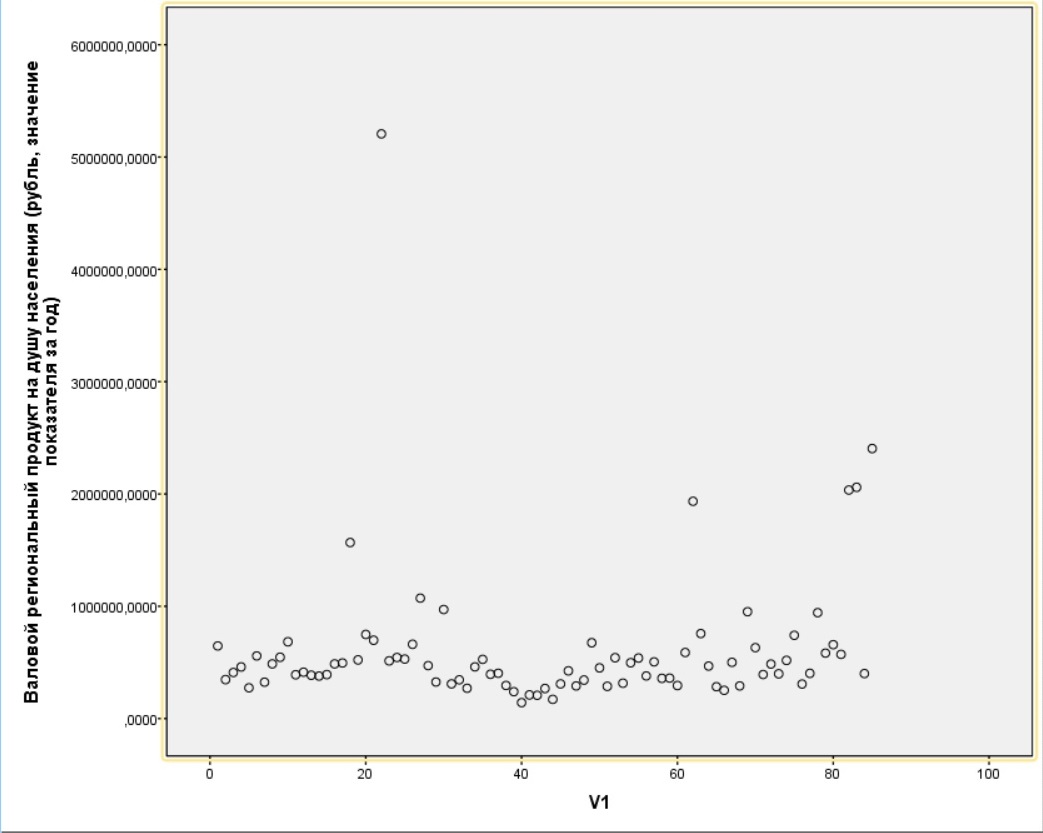
Рисунок 1.3.1.4 отображает точечный график для средимесячной заработной платы, данные график выглядит уже интереснее предыдущих поскольку распределение значительно отличается. Можно сказать, что большая часть наблюдений находиться в приделах от 25000 до 75000. Подозрительными на выбросы могут являться около 6 значений, которые находятся выше 75000.

 Рисунок 1.3.1.4 – Точечный график среднемесячная номинальная заработная плата по полному кругу организации

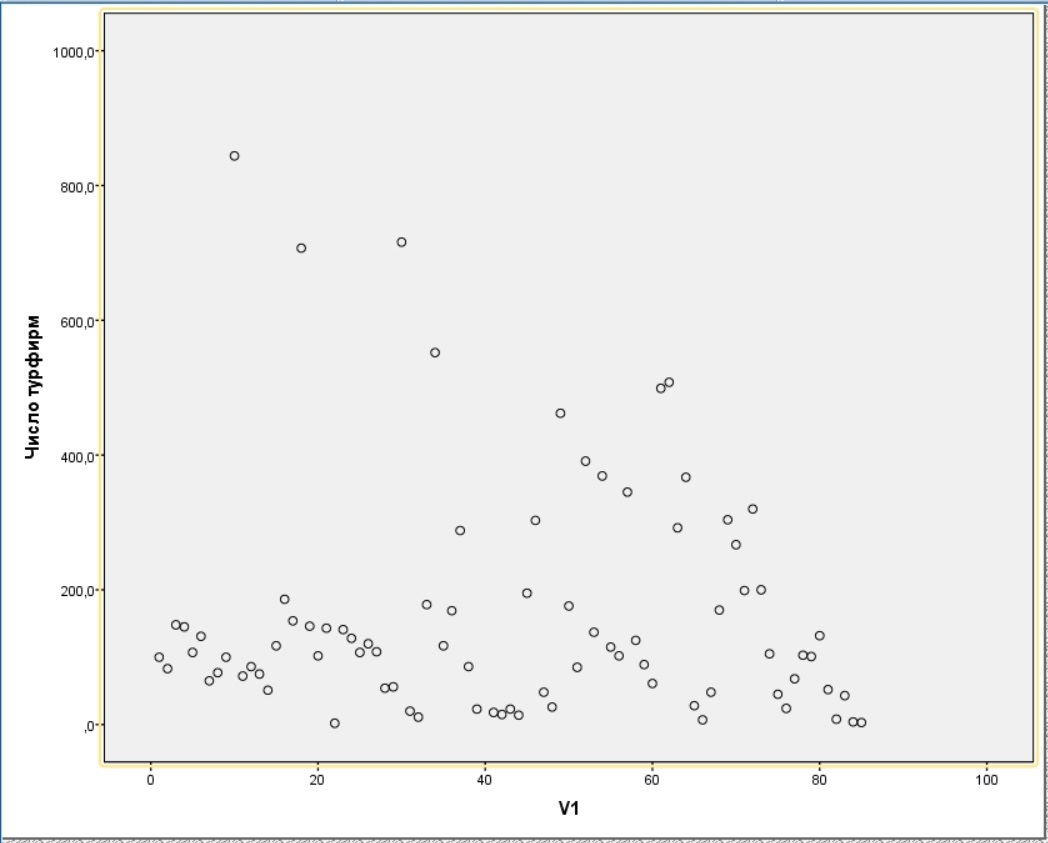
На графике точечного распределения численности размещенных лиц в коллективных средствах размещения (рис. 1.3.1.5) встретилась знакомая картина, где выбиваются 4 региона, 1 из них скорее всего будет точно выбросом, остальные значения расположились в интервале от 0 до 500000.

 Рисунок 1.3.1.5 – Точечный график численность размещенных лиц в коллективных средствах размещения

Посмотрим на рисунок 1.3.1.6, которые отображает точечное распределение для валового регионального продукта, он похожий на некоторые ранее представленные графики тем что один и тот же регион может являться выбросом, есть также еще 4 региона подозреваемых на выброс, при этом ранее такими регионами были те, что находятся в интервале от 10 до 30, а сейчас это регионы с номерами от 60 до 90. Валовой региональный продукт чаще всего принимает значения в приделах от 0 до 2000000.

 Рисунок 1.3.1.6 – Точечный график валовой региональный продукт на душу населения

На рисунке 1.3.1.7 изображен график для числа турфирм в регионах РФ. Он отличается от остальных больше всего, я бы сказала, что в основном показатели расположились в интервале от 0 до 200 единиц, сложно сказать могут ли быть значения от 200 до 600 выбросами, но значения больше чем 600 вполне сойдут за выброс.

 Рисунок 1.3.1.7 – Точечный график число турфирм

* + 1. Листовая диаграмма

Листовая диаграмма показывает больше информации, суммируя форму набора данных (распределение) и предоставляя дополнительные сведения об отдельных значениях. Использование этой формы анализа данных необходимо, чтобы находить медианы, определять итоги и определять режимы наборов данных, обеспечивая ценную информацию о тенденциях и закономерностях в больших наборах данных. Сразу же можно отметить, что если мода будет находиться ближе к 1 или 4 квартилю, то распределение очевидно не будет симметрично, симметрию можно получить, если мода будет близка ко второму квартилю.

Таким образом по рисунку 1.3.2.1 можно сказать, что вероятно мода находится там, где частота 49, также экстремальным значением является оценка туристического потока у Владимирской области и те регионы, у которых принимают значение больше этого.

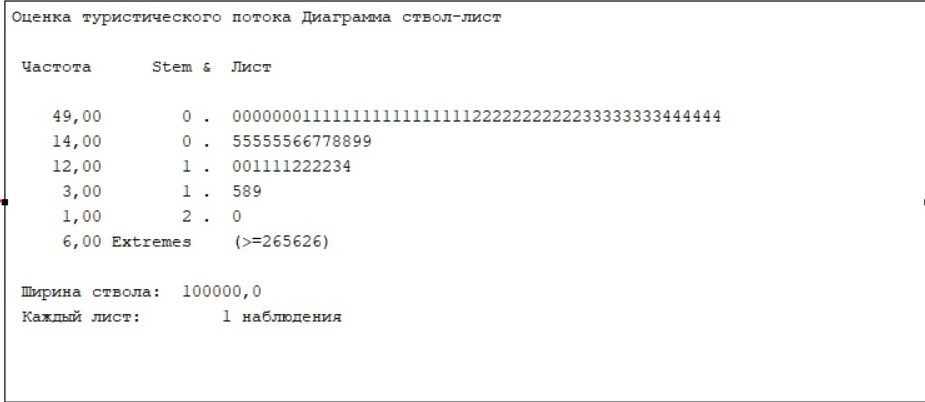
 Рисунок 1.3.2.1 – Диаграмма стебель-листья для оценки туристического потока

Диаграмма стебель-листья для второй переменной похожа по своей структуре на оценку туристического потока, так как мода также находится где-то в первом стебле, а значит расположена ближе к первому квартилю. (рис.1.3.2.2). Экстремальное значение принимает Свердловская область и остальные области, у которых показатель больше данного.

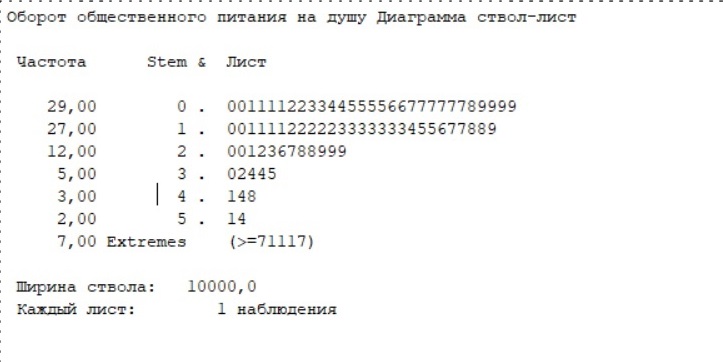
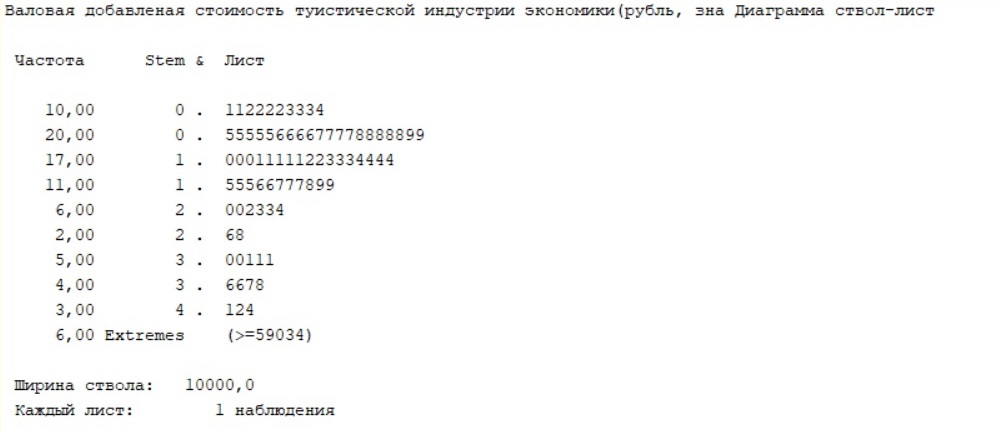
 Рисунок 1.3.2.2 – Диаграмма стебель-листья оборот общественного питания на душу

Рисунок 1.3.2.3 отображает стебль-листья для валовой добавленной стоимости туристической индустрии экономики, видно, что мода находится во 2 стебле, а первым экстремальным значением является республика Татарстан.

 Рисунок 1.3.2.3 – Диаграмма стебель-листья валовой добавленной стоимости туристической индустрии экономики

Рассмотрим листовую диаграмму для среднемесячной номинальной заработной платы по всему кругу организаций. (рис.1.3.2.4). Мода попадает во второй стебель, значит, будет находиться чуть дальше первого квартиля. Нижней границей для экстремальных выбросов является показатель Мурманской области.

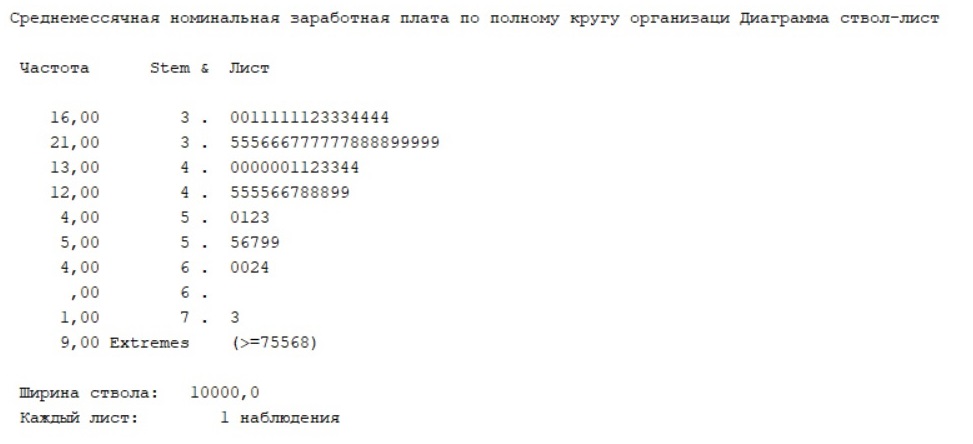
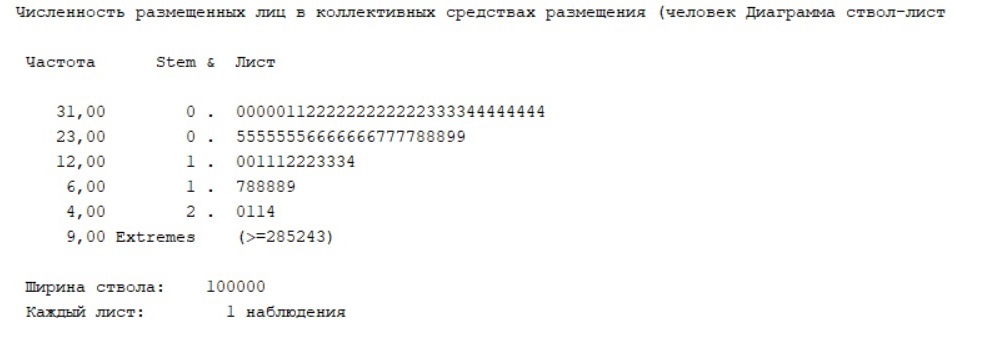
 Рисунок 1.3.2.4 – Диаграмма стебель-листья среднемесячная номинальная заработная плата по полному кругу организаций

Рисунок 1.3.2.5 показывает листовую диаграмму для численности размещенных лиц в коллективных средствах размещения, мода оказалась в первом стебле и скорее всего будет находиться близко к первому квартилю, экстремальным выбросом очевидно является показатель республики Башкортостан.

 Рисунок 1.3.2.5 – Диаграмма стебель-листья по численности размещенных лиц в коллективных средствах размещений

Рассмотрим диаграмму для валового регионального продукта(рис.1.3.2.6). Мода оказалась в третьем стебле диаграммы, скорее всего она будет находиться между первым и вторым квартилем. Экстремальным выбросом можно выявить Краснодарский край.

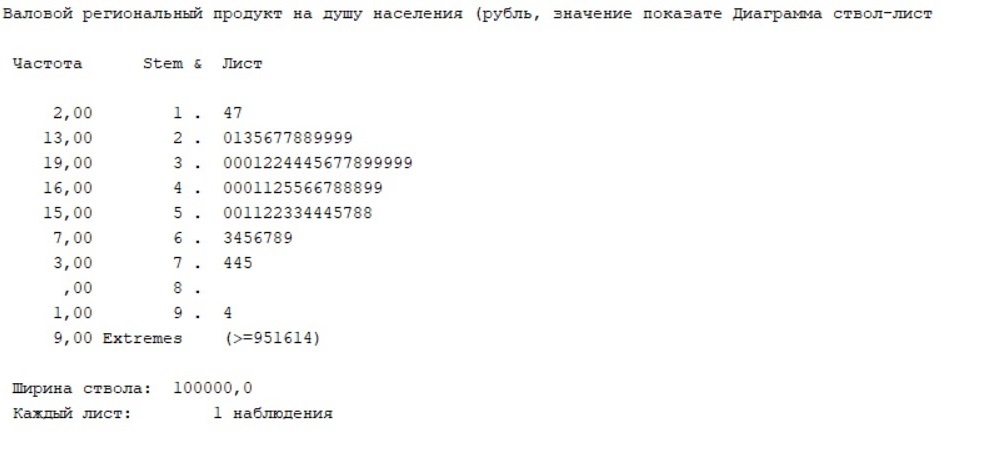
 Рисунок 1.3.2.6 – Диаграмма стебель-листья валового регионального продукта на душу населения

Диаграмма стебель-листья для числа турфирм показывает, что мода находиться близко к первому квартилю, так как она попала в первый стебель, экстремальным значением является Пермский край.

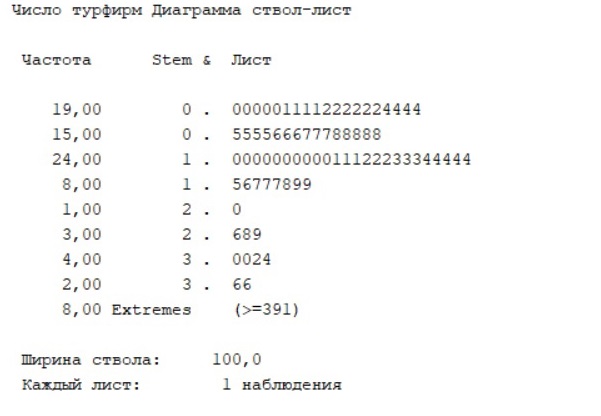
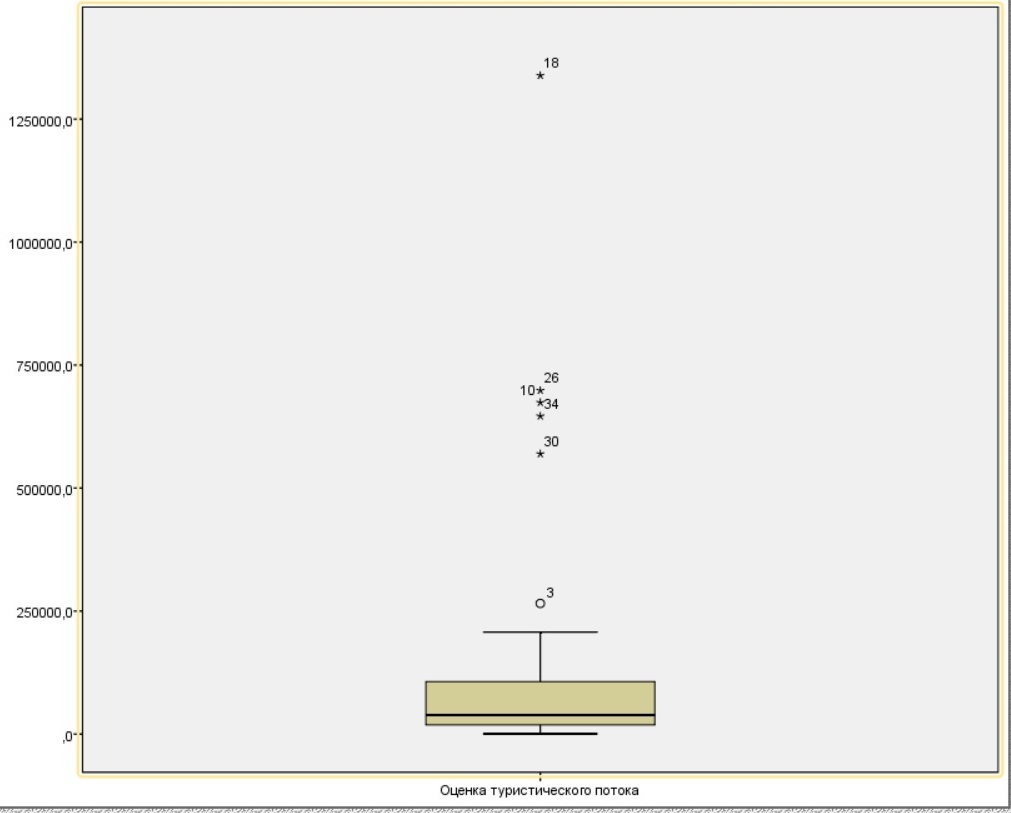


Рисунок 1.3.2.7 – Диаграмма стебель-листья для числа турфирм

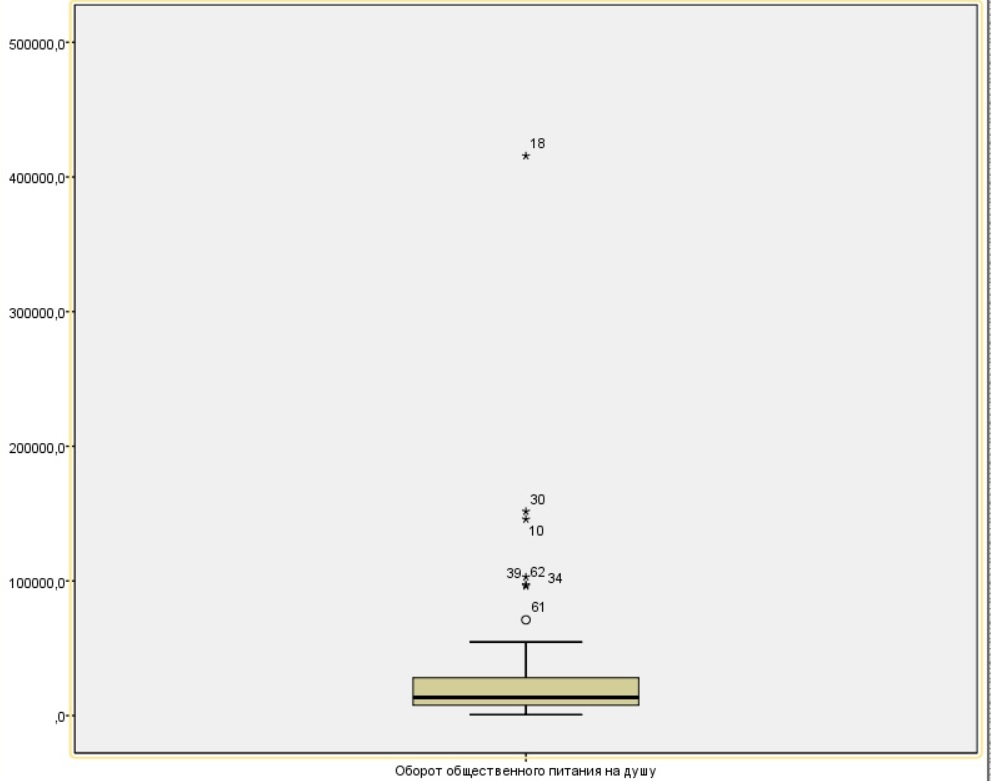
* + 1. Диаграмма выбросов

Диаграмма Box plot («ящик с усами») показывает распределение данных внутри выборки, на ней можно легко найти выбросы, а также расположение медианы.

Рассмотрим рисунок 1.3.3.1, на котором представлен ящик с усами для оценки туристического потока, ранее я предполагала по точечному графику, что здесь будет 4 выброса, ящик с усами показал же, что выбросов 6, из них 5 экстремальных выбросов, которые изображены звездочками. Выбросами являются следующие регионами с такими значениями: Владимирская область- 265626, Московская область 673985, республика Коми- 27720, Ленинградская область- 698385, г. Санкт-Петербург- 569559, Краснодарский край- 646352. По виду диаграммы можем сказать следующее: распределение данных несимметрично, медиана смещена ближе к первому квартилю (нижняя граница «ящика с усами»).

 Рисунок 1.3.3.1 – Ящик с усами оценки туристического потока

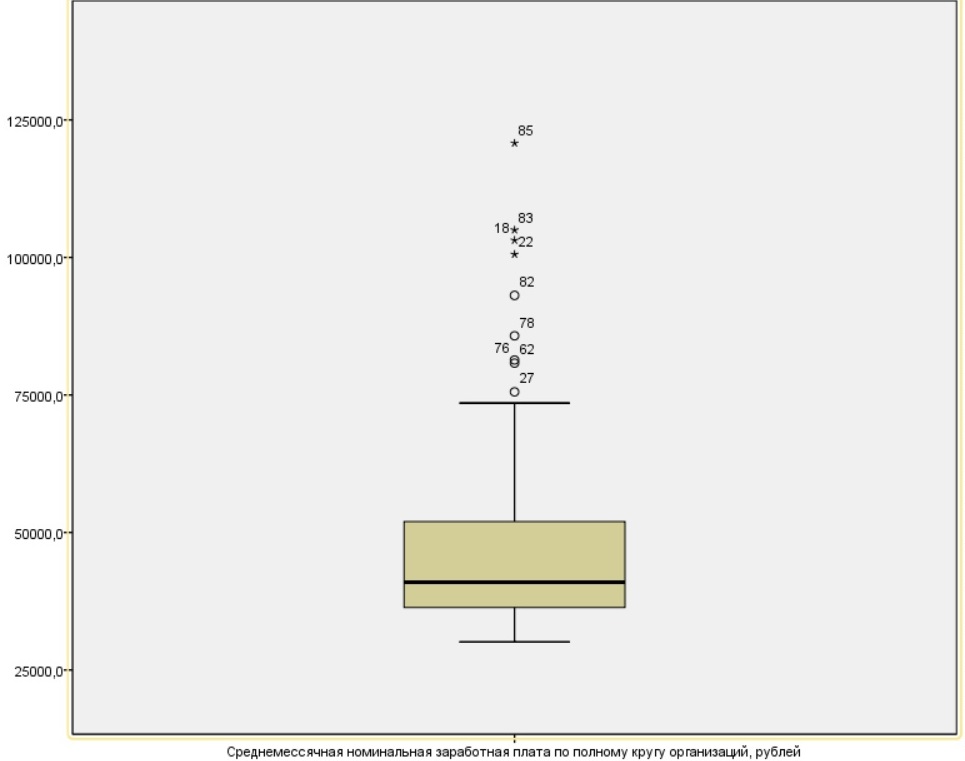
Ящиковая диаграмма для оборота общественного питания имеет 7 выбросов из которых 6 экстремальных, по точечному распределению было выдвинуто предположение, что выбросов будет 4, рассмотрим какие конкретно регионы попали в категорию выбросов и какие у них показатели: Московская область-145863, г. Москва- 415608.2, г. Санкт-Петербург- 151632.3, Краснодарский край-646352, республика Дагестан-96123.2, Свердловская область- 71116.9, Тюменская область-97198.8. Распределение данных также несимметрично, медиана смещена ближе к первому квартилю.

 Рисунок 1.3.3.2 – Ящик с усами для оборота общественного питания на душу

На рисунке 1.3.3.3 мы наблюдаем ящиковую диаграмму для валовой добавленной стоимости туристической индустрии экономики. Предполагалось, что количество выбросов будет равно 4, однако получилось, что их 5 и из них 4 экстремальные, рассмотрим подробно что за области имеют выбросы: Московская область- 175891.8,г.Москва-545097.4,г.Санкт-Петербург- 151234.2,Краснодарский край-153020.6, Свердловская область-59914.2, республика Татарстан- 59034.2. Распределение данных также несимметрично, медиана смещена ближе к первому квартилю.

 Рисунок 1.3.3.3 – Ящик с усами валовая добавленная стоимость туристической индустрии экономики

Рассматривая диаграмму для среднемесячной номинальной заработной платы становиться понятно, что с помощью точечного графика было определено намного меньше выбросов. (рис.1.3.3.4). Ящик с усами показывает, что их тут 9 из них 4 экстремальные, рассмотрим регионы и города, относящиеся к выбросам и их показатели зарплат: г. Москва- 103124, Ненецкий автономный округ-100573,Мурманская область-75568, Тюменская область-80794, республика Сахп-81361, Камчатский край-85759,Магаданская область-93100,Чукотский автономный округ-120769. Распределение данных почти симметрично, медиана находится ближе ко второму квартиле.

 Рисунок 1.3.3.4 – Ящик с усами для среднемесячной номинальной заработной платы по полному кругу организаций

Посмотрим рисунок 1.3.3.5, где представлена ящиковая диаграмма для численности размещенных лиц в коллективных средствах размещений, я предполагала, что будет 4 выброса, но оказалось, что их 9 и 5 из них экстремальные, вот список регионов/городов, попавших в выбросы, а также их показатели: Московская область-907519, г.Москва-2225957,Ленинградская область-337641,г.Санкт-Петербуг-902574,Краснодарский край-1387993, республика Башкорторстан-285243, республика Татарстан-465440, Свердловская область-333964,Тюменская область-392083. Распределение данных несимметрично, медиана смещена ближе к первому квартилю.

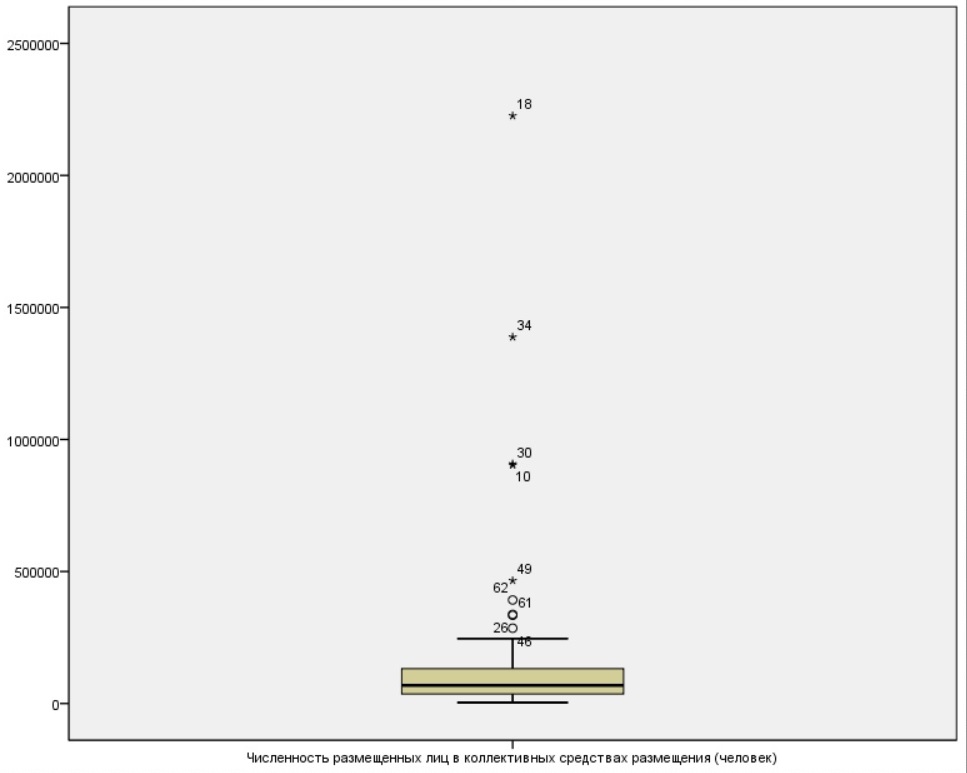
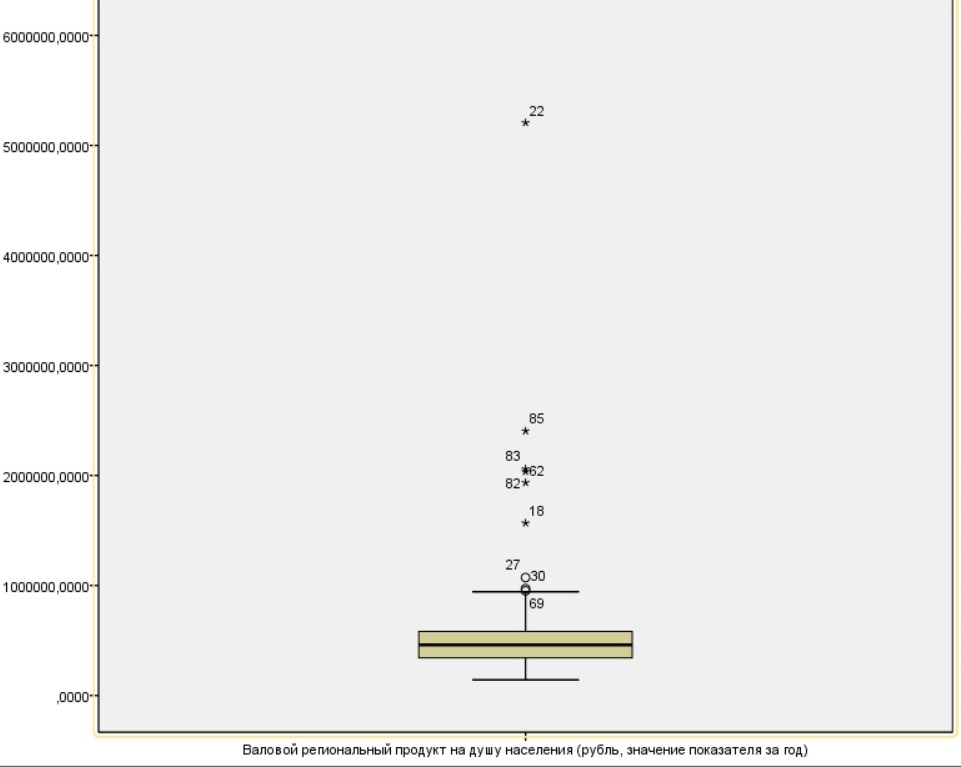
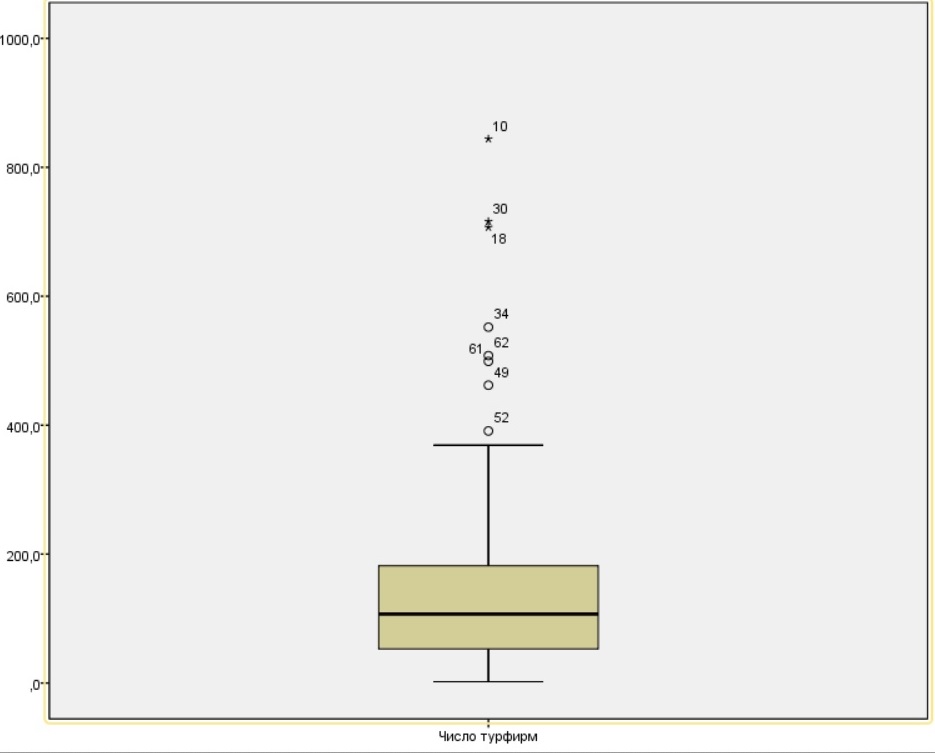
 Рисунок 1.3.3.5 – Ящик с усами численность размещенных лиц в коллективных средствах размещений

Рисунок 1.3.3.6 показывает ящик с усами валового регионального продукта, здесь имеется 9 выбросов и 5 из них экстремальные, по точечному графику было предположено 5 выбросов, рассмотрим их: г.Москва-1567644.8, Ненецкий автономный округ-5206287.1,Мурманская область-1072337.1,Тюменская область-1934463.9,Краснодарский край-951613.7,Магаданская область-2035007,Сахалинская область-2059206.5,Чукотский автономный округ-2404271.2, г.Санкст-Петербург-971158. Распределение данных может быть симметрично, медиана смещена ближе ко второму квартилю.

 Рисунок 1.3.3.6 – Ящик с усами для валового регионального продукта на душу населения

Для количества турфирм также представлен ящик с усами(рис.1.3.37), на нем видно, что имеются 8 выбросов и из них 3 экстремальных. Рассматривая точечный график, я предполагала, что выбросов будет около 4-6. К выбросам относятся: Московская область-844, г.Москва-707, г. Санкт-Петербург- 716, Краснодарский край-552, республика Татарстан-462, Пермский край-391, Свердловская область-499, Тюменская область-508. Распределение данных может быть симметричным поскольку медиана находится почти во втором квартиле.

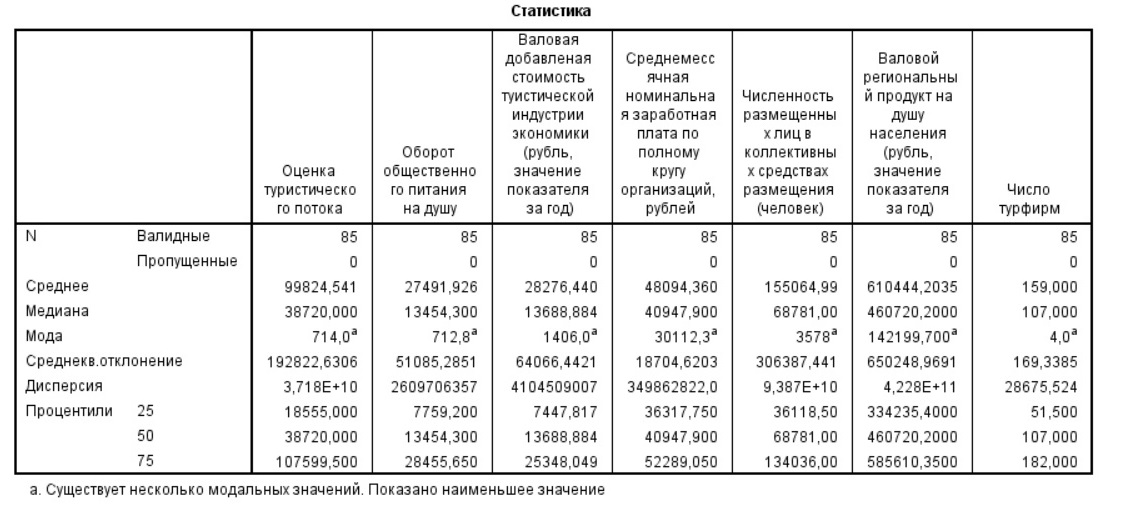
 Рисунок 1.3.3.7 – Ящик с усами для числа турфирм

* 1. Характеристики положения и разброса

Сравнивая между собой моду, медиану и среднее значение можно говорить об асимметрии графиков и распределении данных, рассматривая размах вариации, коэффициент вариации, дисперсию и стандартное отклонение мы наблюдаем за изменчивостью данных.

* + 1. Средняя, медиана и мода

На рисунке 1.4.1.1 представлена сводная таблица со значениями моды, медианы и среднего для всех рассматриваемых переменных. Если значение моды, медианы и среднего равны между собой, то распределение является нормальным и симметричным, в тако распределении может имеется как одна и более мод. Если среднее меньше медианы, а та в свою очередь меньше моды, то наблюдается левостороннее распределение или так называемое отрицательно асимметричное, в ином случае это будет правостороннем распределении (с положительной асимметрией).

 Рисунок 1.4.1.1 – Характеристики положения для всех данных по столбцам

Получилось, что у всех переменных правостороннее распределение, так как среднее значение больше медианы, а медиана больше моды. Можно сказать, что среднемесячная номинальная заработная плата могла бы иметь нормальное распределение без скошенностей.

* + 1. Размах вариации, коэффициент вариации, дисперсия и стандартное отклонение

Размах вариации несет в себе лишь информацию об пределах, в которых изменяется величина признака в изучаемой совокупности. Рассмотрим рисунок 1.4.2.1, на котором показана таблица с размахом вариации для каждой переменой.

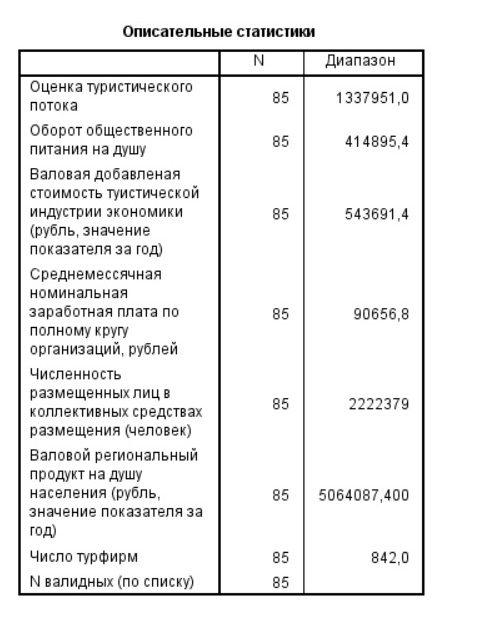


Рисунок 1.4.2.1 – Размах вариации для всех переменных

Основным недостатком данного показателя является то обстоятельство, что максимальные и минимальные значения признака могут быть обусловлены случайными обстоятельствами и в этой связи могут искажать типичный для изучаемой совокупности размах вариации. Таким образом мы имеем размах вариации, который рассчитывался с имеющимися выбросами. Не трудно заметить, что показатель размаха получился очень большим для оценки туристического потока, валового регионального продукта и численности размещенных лиц в коллективных средствах размещения, вероятно это связано с тем, что данные переменные имею больше всего выбросов.

На рисунке 1.4.2.2 приведена таблица, в которой нас интересуют значения дисперсии и стандартного отклонения. Среднее квадратическое отклонение также как и размах вариации отвечает за колебания признака, однако оно показывает, насколько в среднем изменяется величина признака. Дисперсия же представляет из себя квадрат отклонения и имеет зависимость: чем более разбросаны значения в наборе данных, тем выше дисперсия.

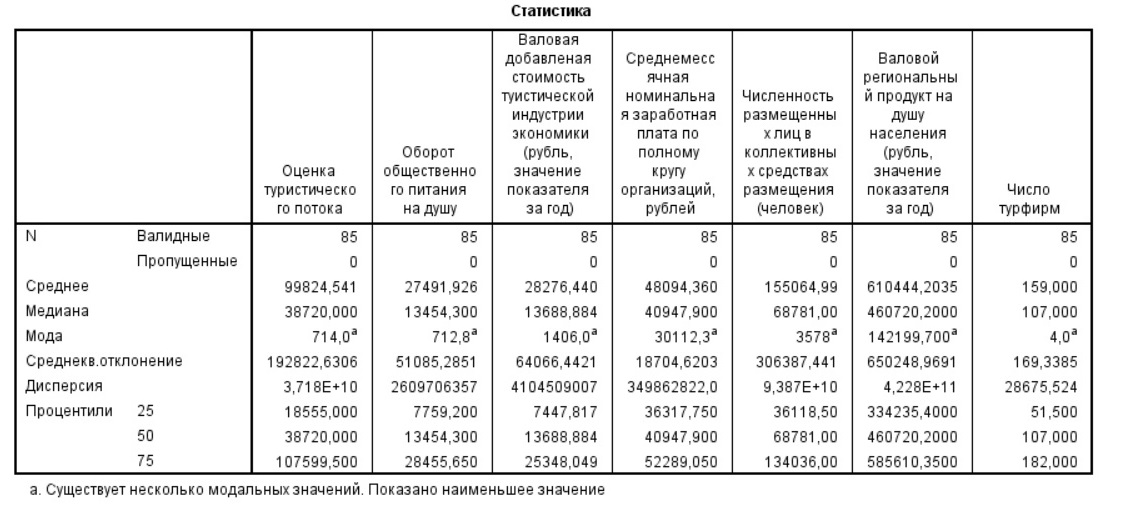


Рисунок 1.4.2.2 – Дисперсия и стандартное отклонение для всех переменных

Как и с размахом вариации выяснилось, что наибольшее значение дисперсии имеют следующие столбцы: оценка туристического потока, численность размещенных лиц в коллективных средствах размещения и валовой региональный продукт. Это переменные, которые больше всего нуждаются в удалении выбросов, поскольку там больше всего экстремальных значений, у других же столбцов показатели тоже огромны, так как имеются выбросы.

Коэффициент вариации — это отношение среднего квадратического отклонения к средней арифметической. По величине коэффициента вариации можно судить о степени вариации признаков совокупности. Чем больше его величина, тем больше разброс значений признаков вокруг средней, тем менее однородна совокупность по своему составу и тем менее представительна средняя. Данный коэффициент был посчитан в SPSS через описательные статистики, где имеется функция отношения. Результаты представлены на рисунке 1.4.2.3.

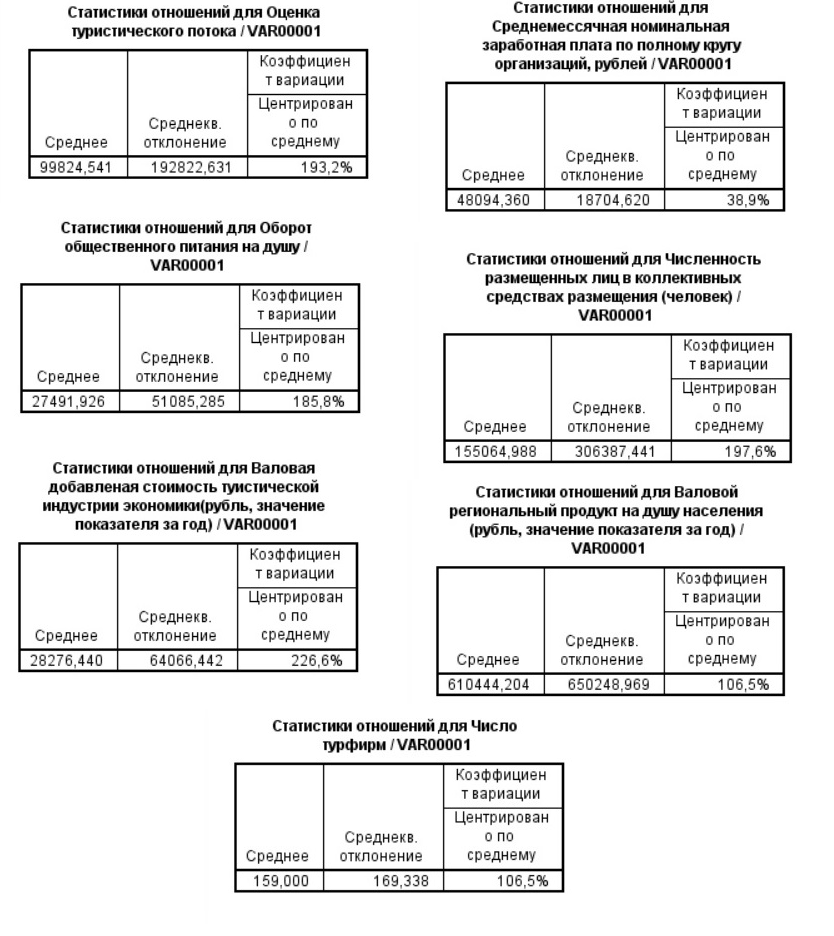


Рисунок 1.4.2.3 – Коэффициент вариации для переменных

Из рассчитанных коэффициентов вариации видно, что у всех значений, кроме среднемесячной номинальной заработной платы по всему кругу организаций имеется огромные разброс значений и это вызывает неоднородность.

* 1. Ранговые характеристики

По аналогии с нахождением медиан в вариационных рядах можно отыскать значение признака у любой по порядку единицы ранжированного ряда. Так, в частности, можно найти значение у единиц, делящих ряд на 4 равные части на 10 и т.п.

* + 1. Квартили и децили

Нулевой квартиль включает в себя самое минимальное значение, 1 квартиль равен 25му перцентилю, 2 квартиль равен медиане и 50-му перцентилю, 3 квартиль — это 75й перцентиль и 4 квартиль равен максимальному числа и 100му перцентилю. На рисунке 1.5.1.1 представлена таблица с первым, вторым и третьим квартилем для всех переменных.

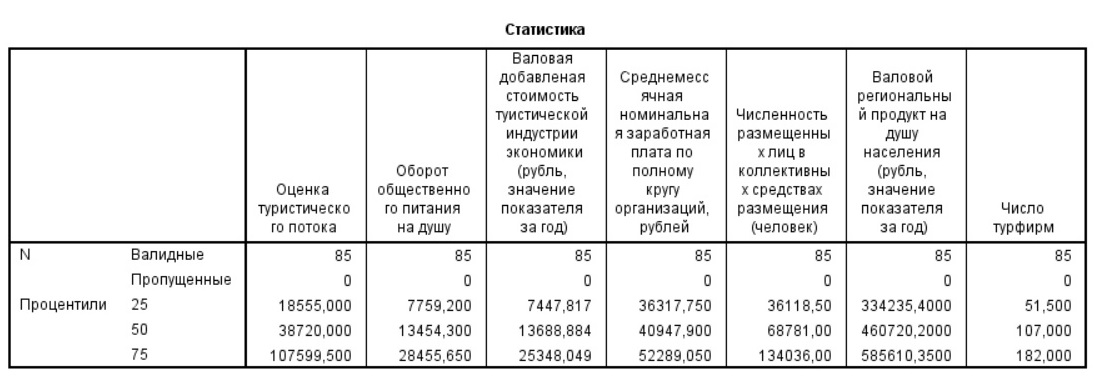


Рисунок 1.5.1.1 – Расчет квартилей

Децилей по своей сущности похожи на квартили, но они делят данные на 10 равных частей, на рисунке 1.5.1.2 представлен расчет децилей в Excel для всех переменных.

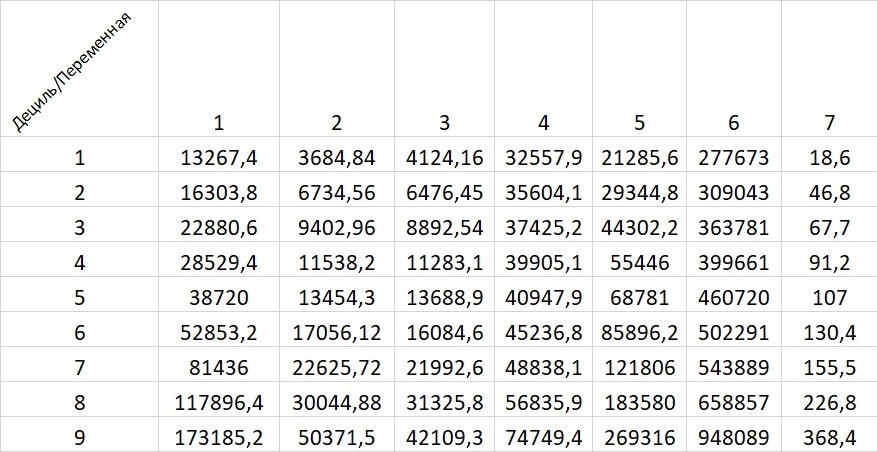


Рисунок 1.5.1.1 – Расчет децилей

* + 1. Межквартильный размах, правила для определения выбросов

В данном подпункте будет проведено исследование на выбросы с помощью расчетов межквартильного размаха и стандартного отклонения. Межквартильный размах был вычислен с помощью калькулятора как , где значение 1 и 3 квартиля нам известны по расчетам из SPSS. Стандартное отклонение было также найдено с помощью функционала SPSS ранее.

Таким образом межквартильный размах для переменных принимает следующие значения: оценка туристического потока-89044.5, оборот общественного питания-20696.45, валовая добавленная стоимость туристической индустрии экономики-17900.232, среднемесячная номинальная зарплата по всему кругу организаций- 15971.3, численность размещенных лиц в коллективных средствах размещения-97917.5, валовой региональный продукт -251374.95, число турфирм-130.5.

* + - 1. 1.5 IQR

Для того чтобы найти выбросы с помощью правила 1.5IQR необходимо вычислить верхнюю и нижнюю границу интервала, если в него будет входить значение рассматриваемой переменной, то это не выброс, иначе мы записываем его в список выбросов. Нижняя граница считается как , а верхняя граница .

Таким образом получаем для каждого рассматриваемого столбца их индивидуальные интервалы. Для оценки туристического потока придельные значения находятся в диапазоне . У оборота общественного питания нормальные значения находятся . Валовая добавленная стоимость туристической индустрии экономики имеет значения, не являющимися выбросами в интервале . Выбросы у среднемесячной номинальной зарплаты по всему кругу организаций располагается вне диапазона.Для численности размещенных лиц в коллективных средствах размещения интервал получился следующий . Диапазон для валового регионального продукта получился таковым. И последний интервал для проверки числа турфирм получился таким .

Сформируем список получившихся выбросов для каждого столбца.

Оценка туристического потока: Владимирская область, Московская область, г. Москва, Ленинградская область, г. Санкт-Петербург, Краснодарский край.

Оборота общественного питания: Московская область, г. Москва, г. Санкт-Петербург, Краснодарский край, республика Дагестан, Свердловская область, Тюменская область.

Валовая добавленная стоимость туристической индустрии экономики: Московская область, г. Москва, г. Санкт-Петербург, Краснодарский край, республика Татарстан, Свердловская область, Тюменская область.

Среднемесячная номинальная зарплата по всему кругу организаций: г. Москва, Ненецкий автономный округ, Тюменская область, республика Саха, Камчатский край, Магаданская область, Сахалинская область, Чукотский автономный округ.

Численность размещенных лиц в коллективных средствах размещения: Московская область, г. Москва, Ленинградская область, г. Санкт-Петербург, Краснодарский край, республика Татарстан, республика Башкортостан, Свердловская область, Тюменская область.

Валовой региональный продукт: г. Москва, Ненецкий автономный округ, Мурманская область, г. Санкт-Петербург, Пермский край, Свердловская область, Тюменская область, Магаданская область, Сахалинская область, Чукотский автономный округ.

Число турфирм: Московская область, г. Москва, г. Санкт-Петербург, Краснодарский край, республика Татарстан, Свердловская область, Тюменская область.

* + - 1. 3 сигмы

Правило трех сигм так же предназначено для обнаружения выбросов, нужно просчитать интервал, где нижняя граница находится по формуле , а верхняя по формуле , где - среднее значение совокупности, а -стандартное отклонение.

Применим это правило и найдем интервалы для всех столбцов.

Для оценки туристического потока придельные значения находятся в диапазоне . У оборота общественного питания нормальные значения находятся . Валовая добавленная стоимость туристической индустрии экономики имеет значения, не являющимися выбросами в интервале . Выбросы у среднемесячной номинальной зарплаты по всему кругу организаций располагается вне диапазона.Для численности размещенных лиц в коллективных средствах размещения интервал получился следующий . Диапазон для валового регионального продукта получился таковым. И последний интервал для проверки числа турфирм получился таким .

Сформируем список получившихся выбросов для каждого столбца.

Оценка туристического потока: Московская область, г. Москва, Краснодарский край.

Оборота общественного питания: г. Москва.

Валовая добавленная стоимость туристической индустрии экономики: г. Москва, Ленинградская область.

Среднемесячная номинальная зарплата по всему кругу организаций:

Численность размещенных лиц в коллективных средствах размещения: г. Москва, Краснодарский край, Сахалинская область, Чукотский автономный округ.

Валовой региональный продукт: Ненецкий автономный округ.

Число турфирм: Московская область, г. Москва, г. Санкт-Петербург.

* 1. Z-преобразование

Z-показатель рассчитывается для каждого значения определенной совокупности по формуле , где х- индивидуальное значение, μ-среднее значение, σ: стандартное отклонение. По вычисленным значениям можно говорить на сколько стандартных отклонений отличается данное значения от среднего.

Так называемый Z столбец строиться автоматически в SPSS, если делать расчет минимума и максимума через функцию «Описательная статистика». На рисунке 1.6.1 представлена таблица максимального и минимального значения Z для каждого столбца.

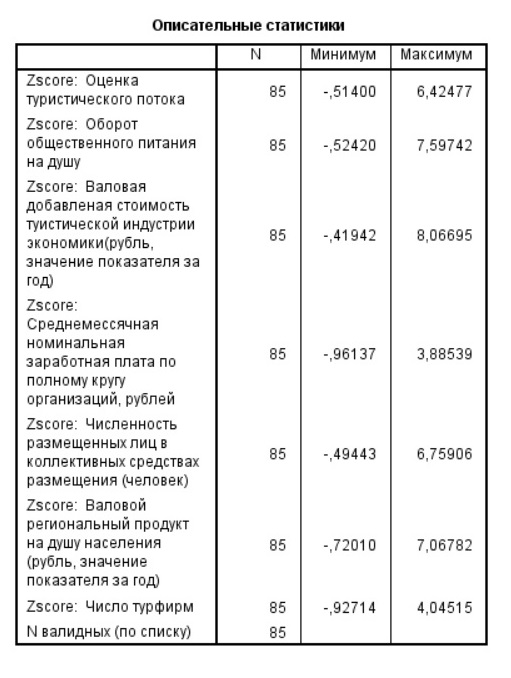


Рисунок 1.6.1 – Максимальное и минимальное значение для Z-столбцов

Как видно в основном значения колеблются от -1 до 6-7. Интерпретировать это можно следующим образом Положительный z-показатель указывает на то, что конкретное значение больше среднего, соответственно отрицательная z-оценка показывает, что значение меньше среднего, а z-оценка, равная нулю дает понять, что значение равно среднему.

Рассмотрим значения Z-оценок для выбросов по каждому столбцу. Выбросы были взяты из пункта 1.5.2.1, где они находились через 1.5IQR.

Оценка туристического потока: Владимирская область со значением 0.85987, Московская область с z-оценкой равной 2.97766, г. Москва со значением 6.42477, Ленинградская область получила показатель 3.10420, г. Санкт-Петербург с z-оценкой 2.43610, Краснодарский край со значением 2.83435.

Оборота общественного питания: Московская область с показателем 2.31713, г. Москва получила значение 7.59742, г. Санкт-Петербургу присваивается z-оценка 2.43006, Краснодарский край имеет оценку равную 1.47572, республика Дагестан получила значение 1.34346, Свердловская область имеет z-оценку равную 0.85396, Тюменская область со значением 1.36452.

Валовая добавленная стоимость туристической индустрии экономики: Московская область с z-оценкой 2.30410, г. Москва и значение 8.06695, г. Санкт-Петербург получил значение 1.91922, для Краснодарского края значение получилось 1.94711, республика Татарстан с оценкой 0.48009, Свердловская область получила значение 0.49383, Тюменская область с оценкой равной 0.50454.

Среднемесячная номинальная зарплата по всему кругу организаций: г. Москва равно 2.94202, Ненецкий автономный округ со значением 2.80566, Тюменская область с z-оценкой 1.74819, республика Саха получила значение 1.77855, Камчатский край со значением 2.01368, Магаданская область с оценкой равной 2.40613, Сахалинская область с оценкой равной 3.04172, Чукотский автономный округ получил оценку равную 3.88539.

Численность размещенных лиц в коллективных средствах размещения: Московская область получила оценку равную 2.45589, г. Москва с оценкой равной 6.75906, Ленинградская область с показателем оценки 0.59590, г. Санкт-Петербург с z-оценкой 2.43975, Краснодарский край получил оценку равную 4.02408, республика Татарстан по подсчетам получил значение 1.01301, республика Башкортостан со значением 0.42488, Свердловская область с показателем 0.58390, Тюменская область с оценкой 0.77359.

Валовой региональный продукт: г. Москва с оценкой 1.47205, Ненецкий автономный округ после расчетов подучил значение 7.06782, Мурманская область с показателем 0.71033, г. Санкт-Петербург с показателем 0.55473, Пермский край с получившейся оценкой -0.10537, Свердловская область с показателем -0.03410, Тюменская область с показателем 2.03617, Магаданская область со значением 2.19080, Сахалинская область с оценкой 2.22801, Чукотский автономный округ с показателем 2.75868.

Число турфирм: Московская область с показателем 4.04515, г. Москва с оценкой 3.23612, г. Санкт-Петербург с показателем 3.28927, Краснодарский край с оценкой 2.32080, республика Татарстан со значением 1.78932, Свердловская область с z-оценкой 2.00781, Тюменская область с показателем 2.06096.

Таким образом в основном все выбросы имеют значения больше среднего. Это можно подтвердить и диаграммы ящик с усами, которые были представлены ранее.

* 1. Выводы

В результате проделанного предварительного анализа было выявлено много выбросов в данных, к этому выводу можно прийти, опираясь на графическое представление данных, а также на правила нахождения выбросов. Также рассматривались распределения данных, в ходе чего, стало понятно, что данные неоднородные, имеют несколько мод и не относятся к нормальному распределению, кроме того, размах вариации у представленных данных велик. Из-за того, что данные не имеют нормального распределения, стало ясно, что у всех переменных будет правосторонняя скошенность, кроме среднемесячной номинальная заработной платы.

В дальнейшем выбросами будут считаться те данные, которые были выведены в отдельный список при реализации правила 1.5IQR.

# **Глава 2 Корреляционный анализ**

Во второй главе представлен корреляционный анализ, который включает в себя построение корреляционных полей и определение связи между переменными для всех данных, а также для данных с удаленными выбросами, кроме того, проделано построение матриц и расчеты различных коэффициентов корреляции.

1. 1. Построение полей корреляции

На рисунках представленных ниже изображены корреляционные облака, где в качестве оси Y рассматривается оценка туристического потока, а все остальные столбцы рассматриваются как X.

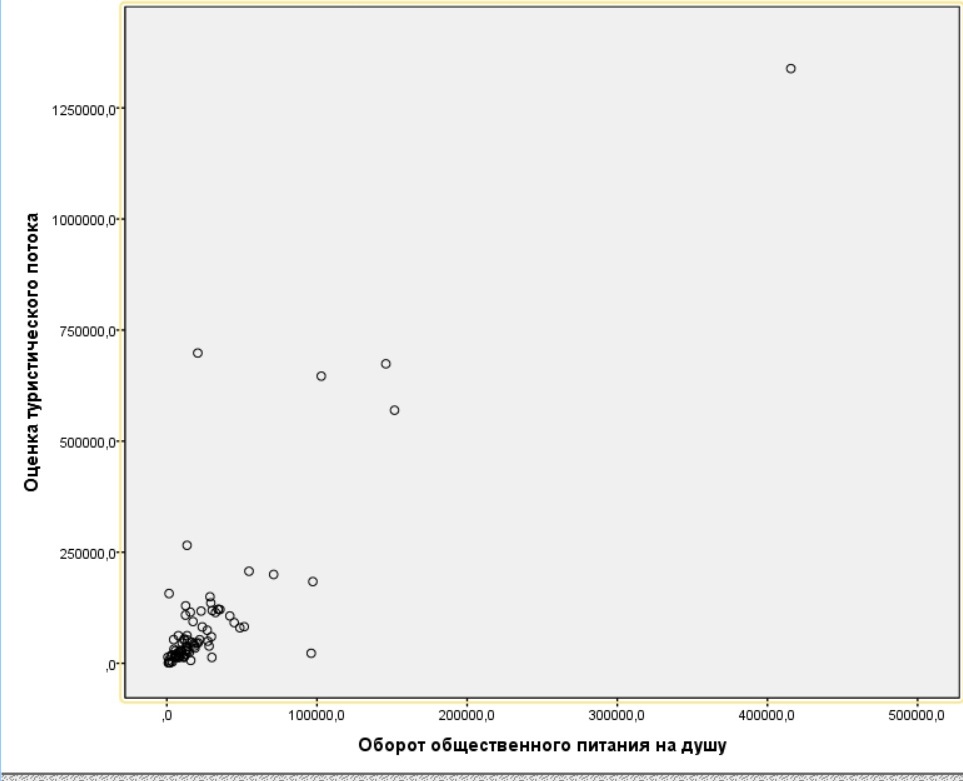


Рисунок 2.1.1 – Корреляционное поле оценки туристического потока и оборота общественного питания

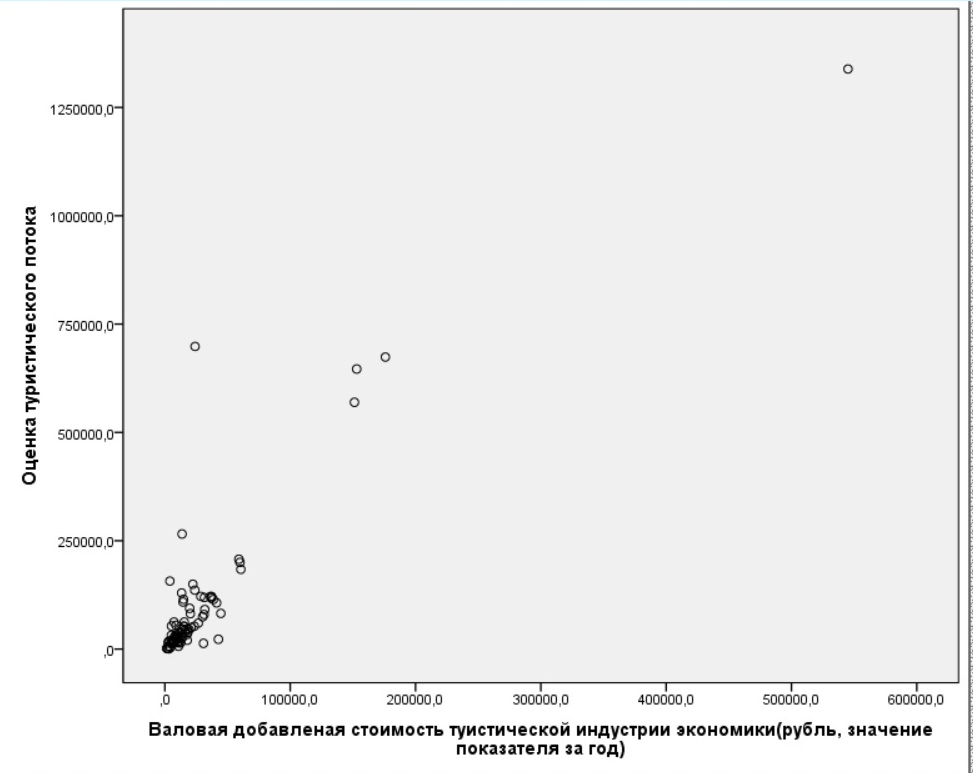
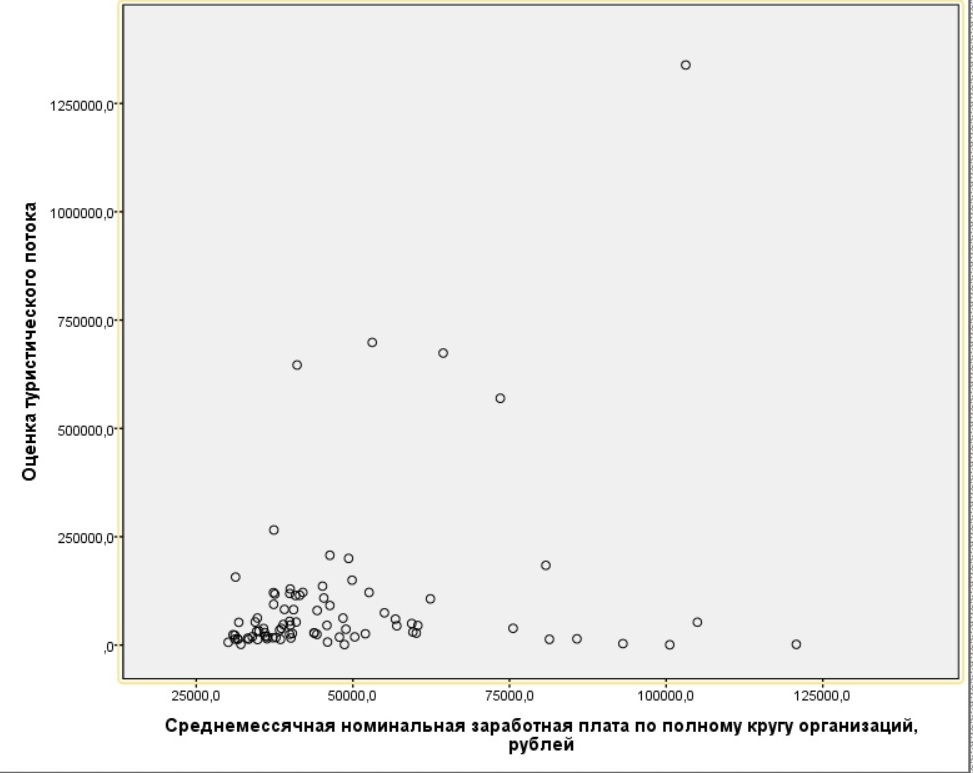


Рисунок 2.1.2 – Корреляционное поле оценки туристического потока и валовой добавленной стоимости туристической индустрии экономики

 Рисунок 2.1.3 – Корреляционное поле оценки туристического потока и среднемесячная номинальная заработная плата по полному кругу организаций

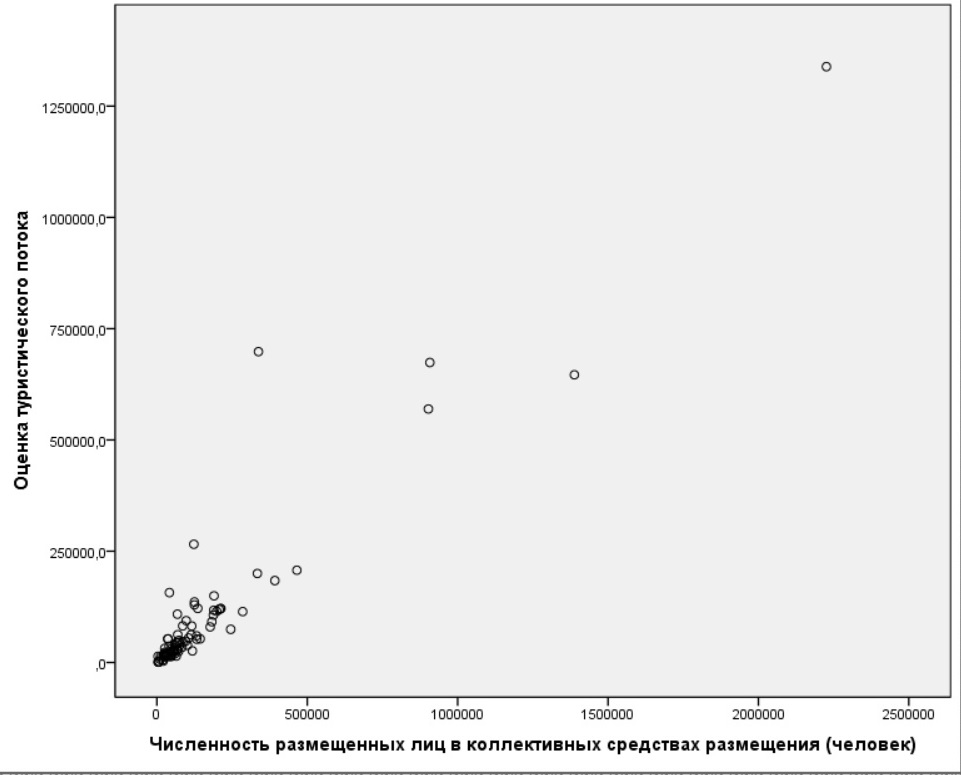
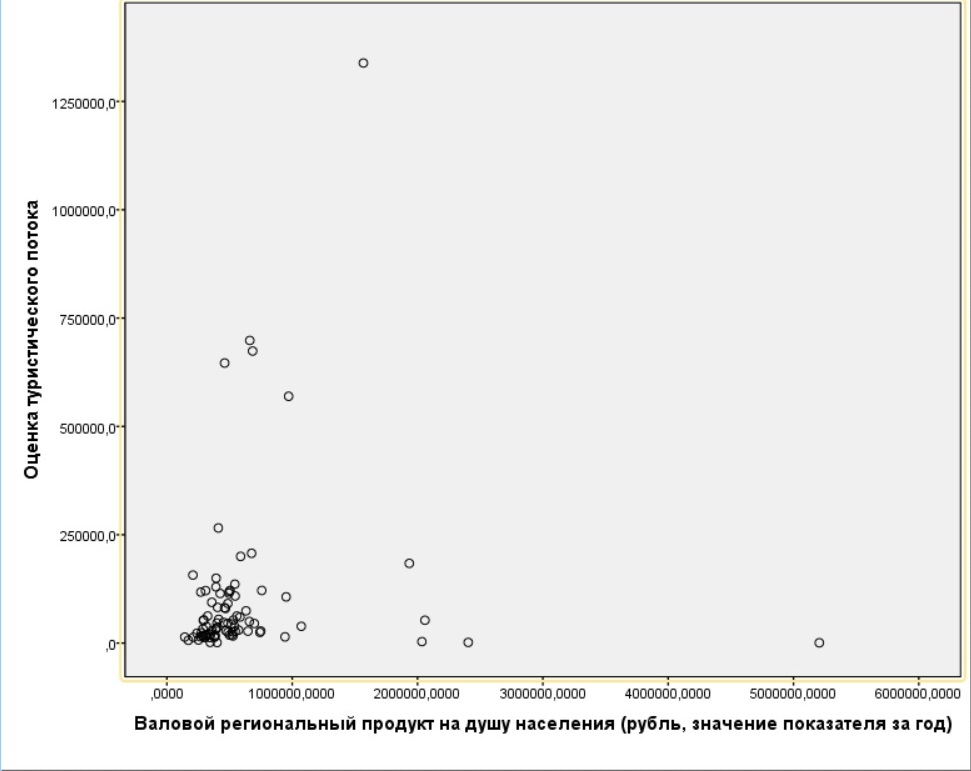
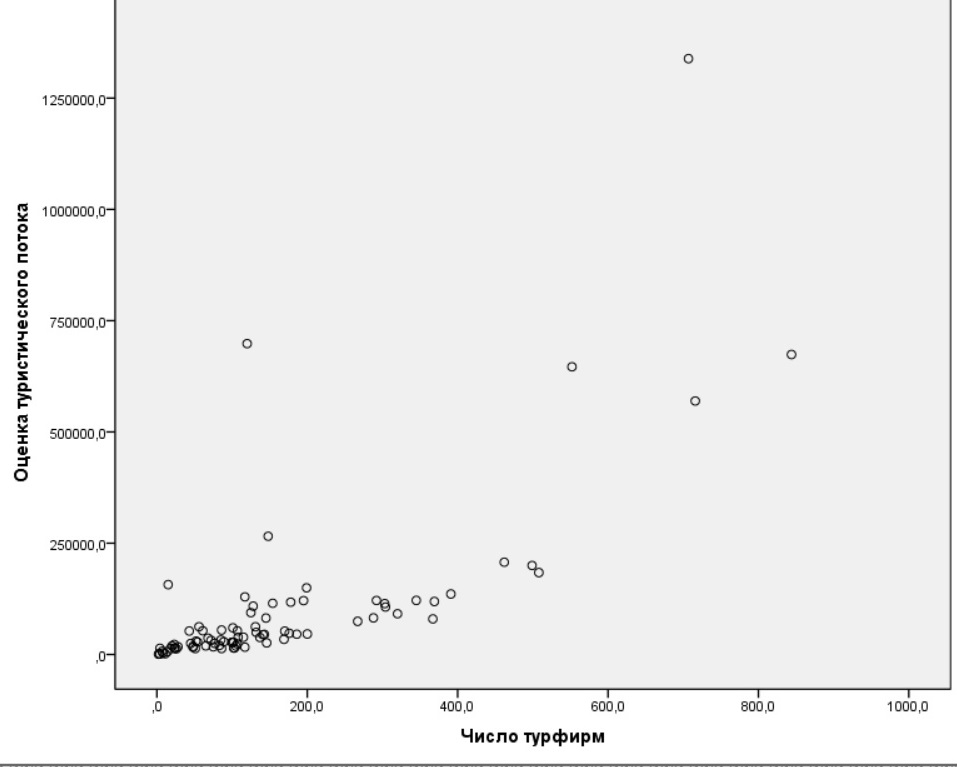


Рисунок 2.1.4 – Корреляционное поле оценки туристического потока и численности размещенных лиц в коллективных средствах размещения

 Рисунок 2.1.5 – Корреляционное поле оценки туристического потока и валового регионального продукта

 Рисунок 2.1.6 – Корреляционное поле оценки туристического потока и числа турфирм

* 1. Описание связи между переменными

Для того чтобы описать связь между переменными необходимо проанализировать рассмотреть силу и направление связи, а также аномальные наблюдения.

По наклону в точечном графике можно судить об направлении связи (направление определяет знак в коэффициенте корреляции), а по ширине об силе и ее тесноте, т.е. насколько тесно расположены точки-объекты около линии регрессии - чем ближе точки к линии, тем сильнее связь.

Рассмотрим точечный график зависимости оценки туристического потока и оборот общественного питания, наблюдается положительная корреляция с прямым направлением, теснота связи, если судить по расположению точек, может быть сильной, если убрать аномальные наблюдения, а именно значения оценки туристического потока, которые больше 500000 и значения оборота общественного питания, которые больше 100000. Из этого можно сделать вывод, что, если при вычислении коэффициента корреляции, то он должен быть положительным и по величине до 0.7.

Точечный график зависимости между оценкой туристического потока и валовой добавленной стоимостью на туристическую индустрии экономики показывает, что корреляция является положительной, связь между показателями прямая, а ее теснота сильная у значений, где показатель оценки турпотока и валовой добавленной стоимости меньше соответственно 500000 и 100000, дальше связь становиться слабой, поэтому модно предположить, что там находятся аномальные значения. Коэффициент корреляции должен быть положительным и находится в диапазоне до 0.9,

Корреляционное облако между оценкой турпотока и среднемесячной номинальной заработной платы по всему кругу организаций выглядит уже интереснее, поскольку наблюдается либо отсутствие связи, либо связь очень слабая, при этом есть значения, которые кучкуются и находятся рядом, будто бы теснота связи сильная, сложно судить о направлении связи. Скорее всего аномальными наблюдениями являются те значения, которые у каждого из показателей превышают соответственно 250000 и 75000. Из этого можно предположить, что коэффициент корреляции будет близок к нулю.

Точечный график оценки туристического потока и численность размещенных лиц в средствах коллективного размещения представляет из себя положительную корреляцию, связь и теснота сильные, однако, если рассматривать значения, которые превышают 500000 у обоих рассматриваемых переменных, то можно сказать, что связь слабая, вероятно там находятся аномальные значения. Коэффициент корреляции должен быть положительным и достигать значения до 1.

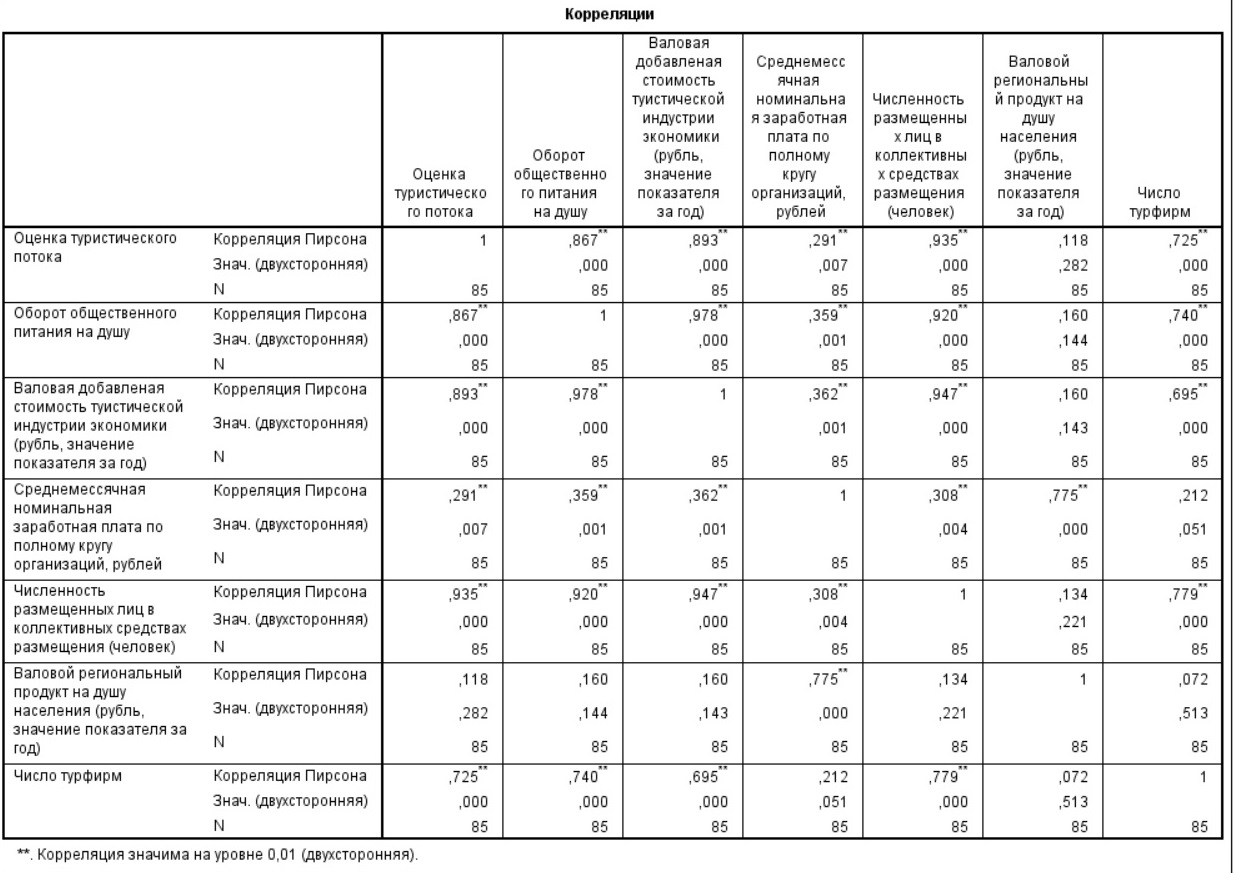
Облако корреляции показателей оценки туристического потока и валового регионального продукта представляет из себя отсутствие связи или же слабую связь между значениями, теснота между объектами сильная в приделах от 0 до 250000 для оценки турпотока и в интервале от 0 до 1000000 для валового регионального продукта, остальные значения имеют слабую тесноту связи и являются аномальными значениями, направление связи либо отсутствует, либо чуть схоже на прямое. Коэффициент корреляции должен быть положительным и близким к 0.

Точечный график для значений оценки турпотока и числа турфирм показывает, что между этими показателями может существовать сильная положительная связь, у которой теснота так же большая. Аномальными наблюдениями можно считать те, что превышают 75000 у оценки туристического потока и 600 у числа турфирм. Коэффициент корреляции должен быть положительным и принимать значение в приделах от 0.6 до 0.8.

* 1. Матрица парных коэффициентов корреляции

Корреляция имеет значение от -1 до 1, где: -1 указывает на совершенно отрицательную линейную корреляцию между двумя переменными,0 указывает на отсутствие линейной корреляции между двумя переменными, 1 указывает на совершенно положительную линейную корреляцию между двумя переменными.

Рассмотрим матрицу парных коэффициентов корреляции(рис.2.3.1). Не имеется отрицательной линейной связи, есть значения близкие к нулю, а именно 0.118, 0.072, 0.164, 0.134, это можно интерпретировать как отсутствие связи между показателями, конкретно получилось, что почти все переменные никак не зависят от валового регионального продукта. У остальных показателей в основном наблюдаются коэффициенты в диапазоне от 0.6 до 0.8, это говорит о положительной линейной корреляции и сильной связи между переменными.

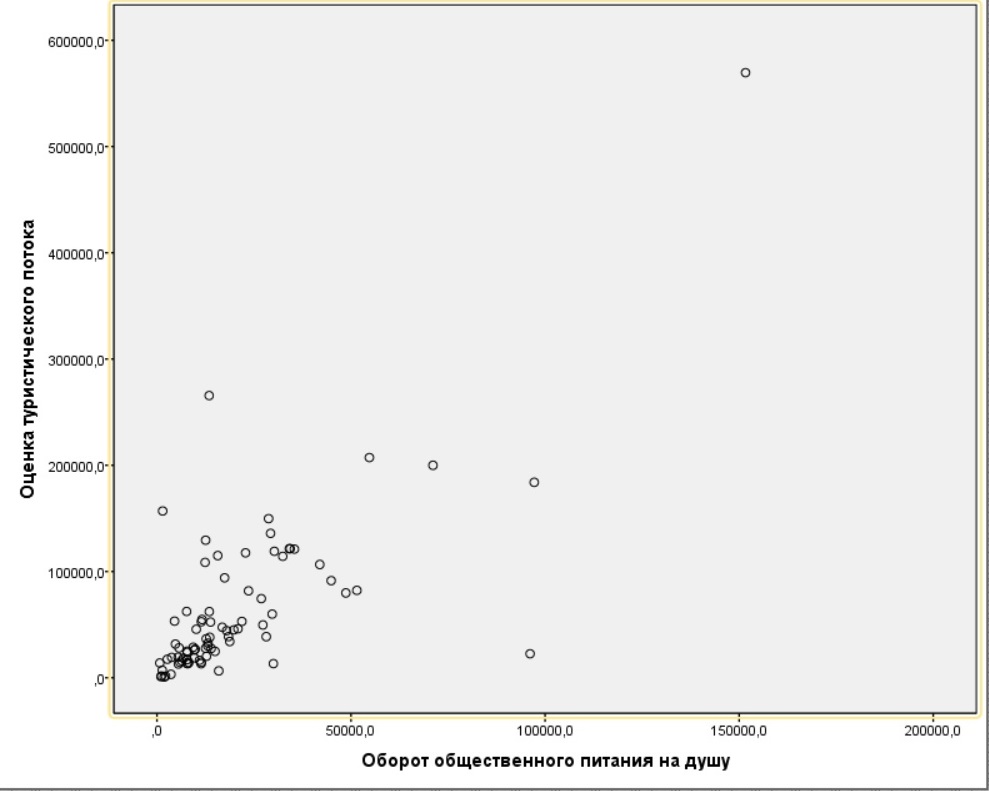


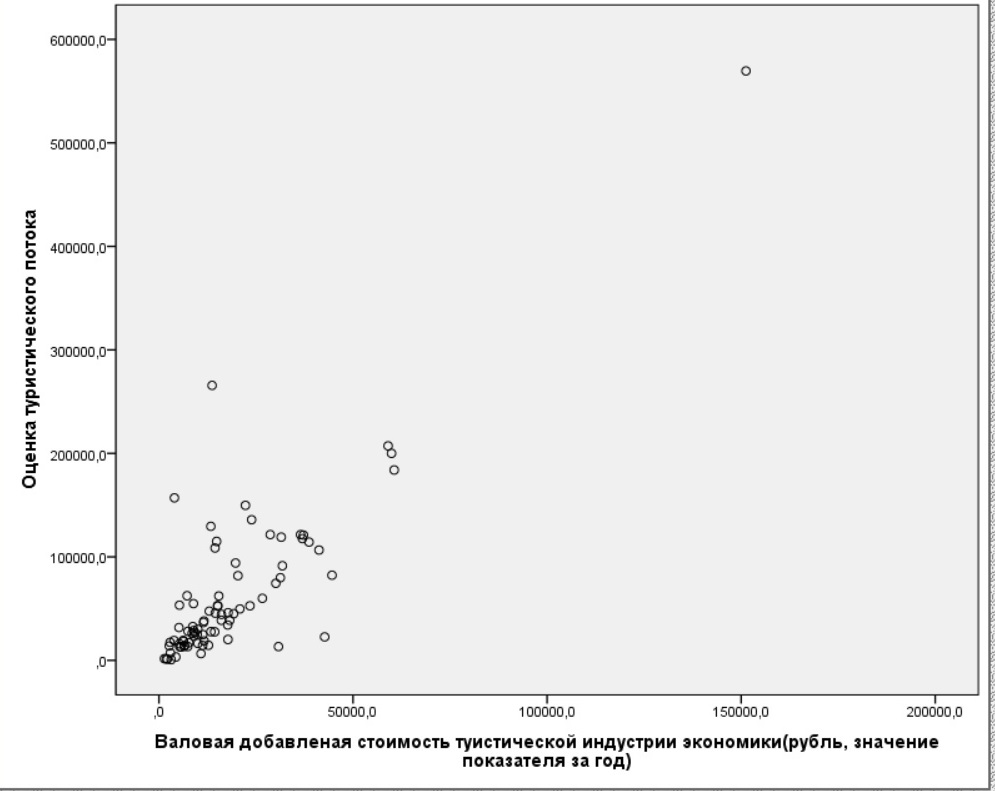
***Рисунок 2.3.1 –Матрица парных коэффициентов корреляции***

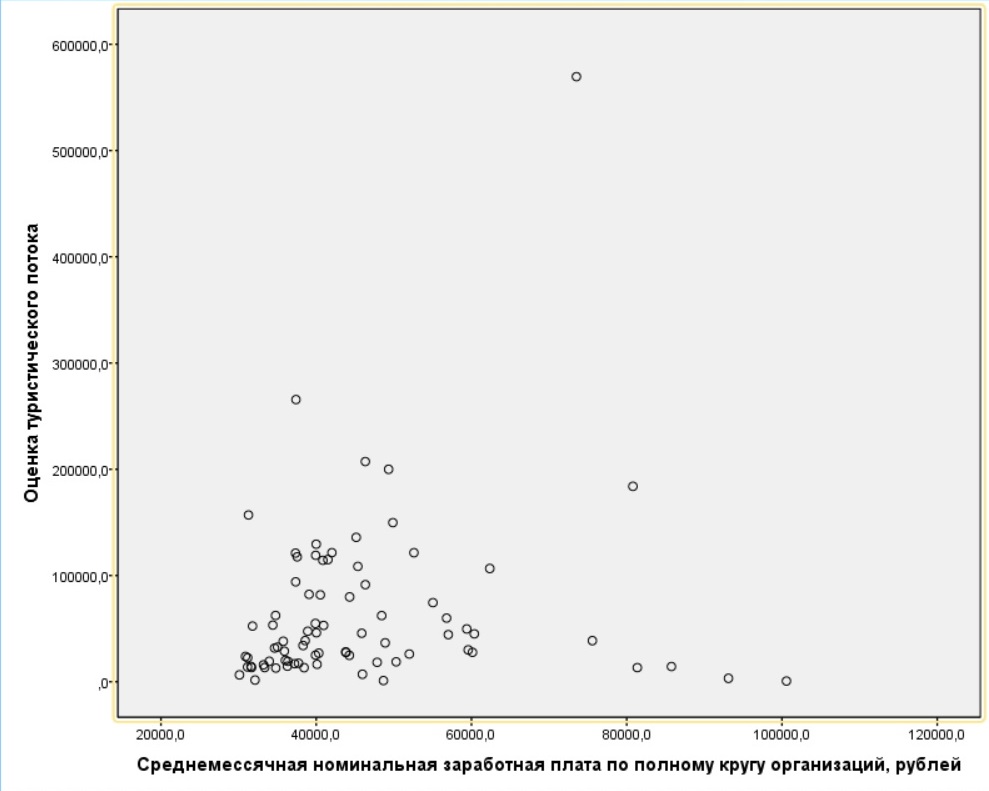
Перечислим по строкам от чего больше всего зависит та или иная переменная. Оценка туристического потока имеет высокие показатели корреляции с переменными: оборот общественного питания, валовая добавленная стоимость туристической индустрии экономики, численность размещенных лиц в коллективных средствах размещения и число турфирм. Оборот общественного питания: оценка туристического потока, валовая добавленная стоимость туристической индустрии экономики, численность размещенных лиц в коллективных средствах размещения и число турфирм. Валовая добавленная стоимость туристической индустрии экономики: оценка туристического потока, оборот общественного питания, численность размещенных лиц в коллективных средствах размещения и число турфирм. Среднемесячная номинальная заработная плата по всему кругу организаций: валовой региональный продукт. Численность размещенных лиц в коллективных средствах размещения: оценка туристического потока, оборот общественного питания, валовая добавленная стоимость туристической индустрии экономики и число турфирм. Число турфирм: оценка туристического потока, оборот общественного питания, численность размещенных лиц в коллективных средствах размещения.

* 1. Построение полей корреляции после удаления выбросов

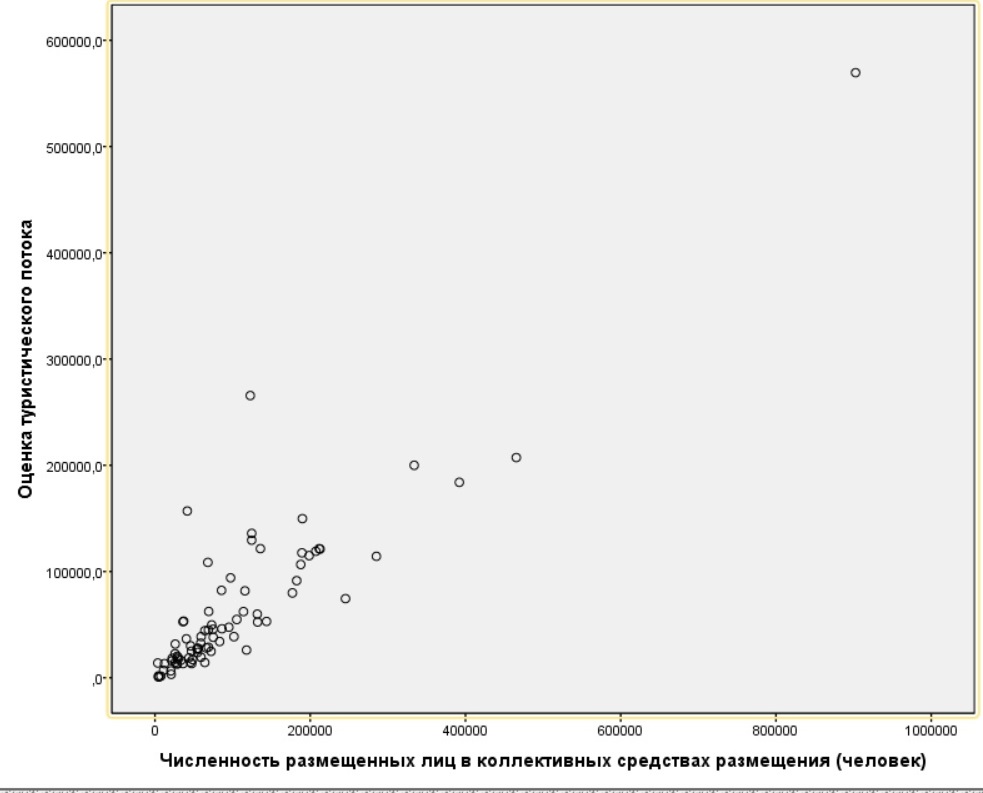
На рисунках представленных ниже изображены корреляционные облака после удаления выбросов, где в качестве оси Y рассматривается оценка туристического потока, а все остальные столбцы рассматриваются как X.

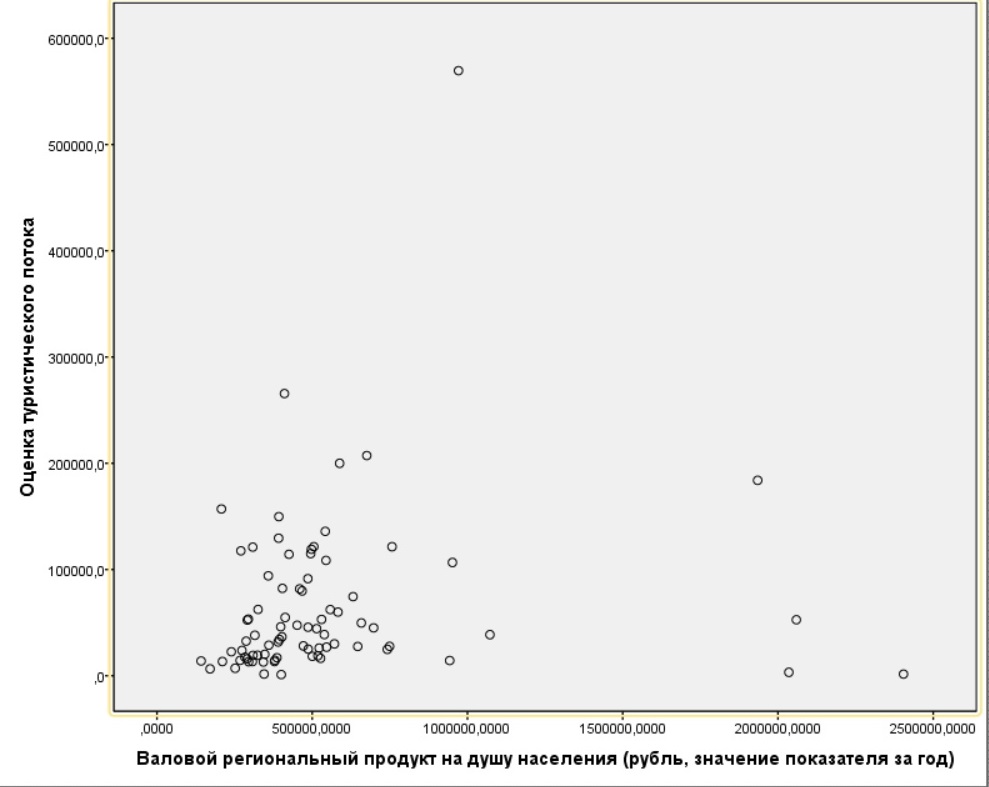
 ***Рисунок 2.4.1 –*** ***Корреляционное поле оценки туристического потока и оборота общественного питания***

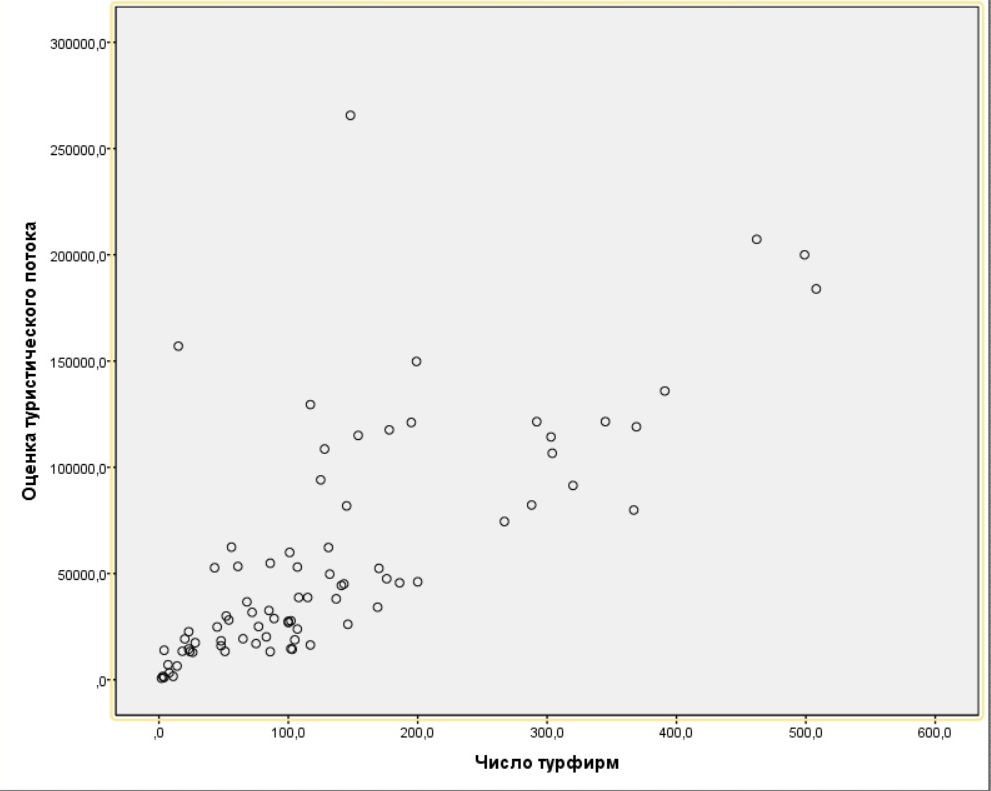
 ***Рисунок 2.4.2 –*** ***Корреляционное поле оценки туристического потока и валовой добавленной стоимости туристической индустрии экономики***



***Рисунок 2.4.3 –*** ***Корреляционное поле оценки туристического потока и среднемесячной номинальной заработной платы***

 ***Рисунок 2.4.4 –*** ***Корреляционное поле оценки туристического потока и численности размещенных лиц в коллективных средствах размещения***

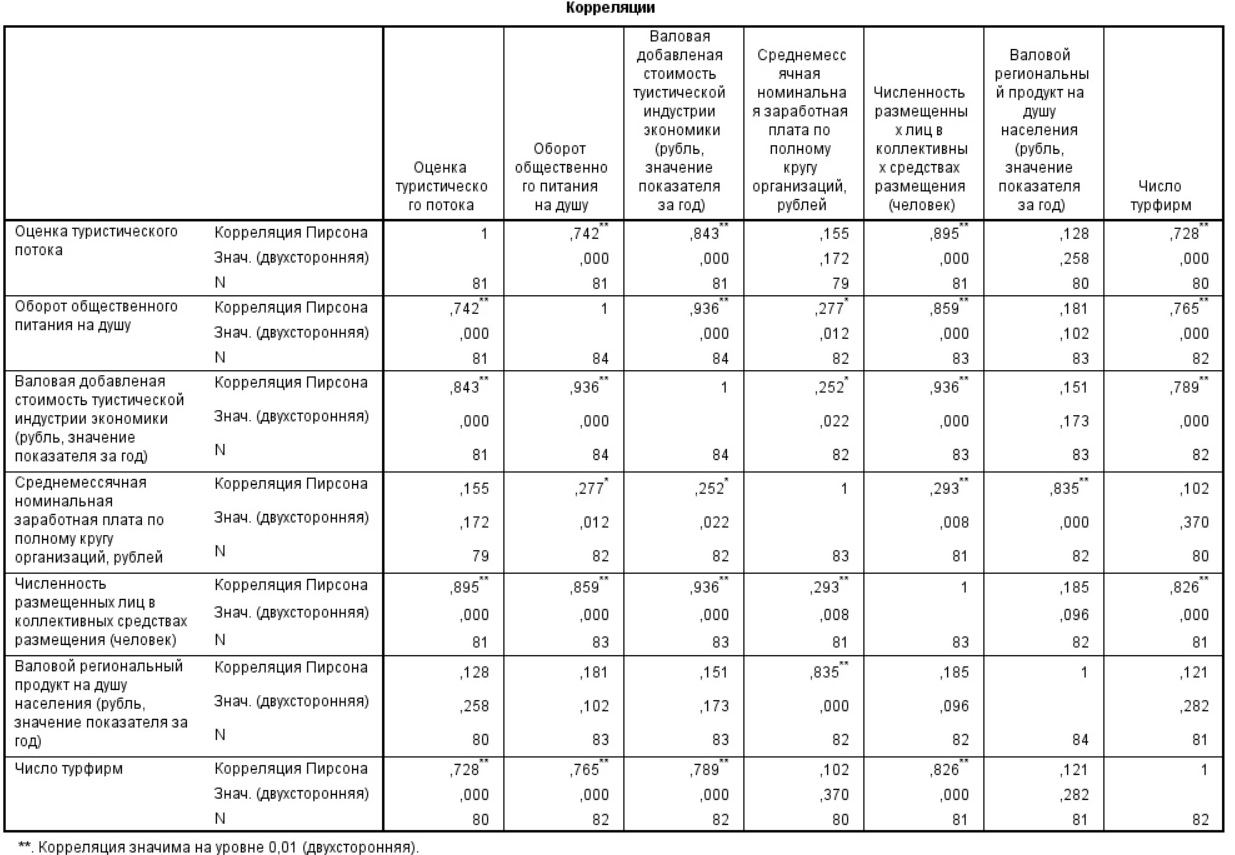
 ***Рисунок 2.4.5 –*** ***Корреляционное поле оценки туристического потока и валового регионального продукта***

 ***Рисунок 2.4.6 –*** ***Корреляционное поле оценки туристического потока и числа турфирм***

Если сравнивать корреляционные поля до и после удаления выбросов, то можно отметить, что из-за того, что поменялся масштаб графиков стала лучше видна теснота между признаками. В основном корреляция между признаками не поменялась, направление связи осталось таким же, но в некоторых случаях появился больший разброс точек.

* 1. Матрица парных коэффициентов корреляции после удаления выбросов

На рисунке 2.5.1 изображена матрица парных коэффициентов корреляции после удаления выбросов. Сравним ее с предыдущей матрицей. В основном все тенденции, которые были описаны ранее остались, все также все коэффициенты, кроме среднемесячной номинальной заработной платы по всему кругу организаций не зависят от валового регионального продукта.

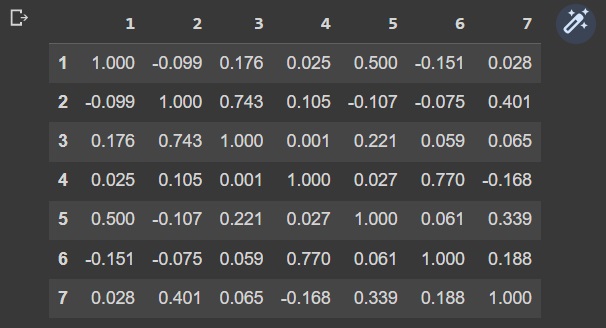
 ***Рисунок 2.5.1 –Матрица парных коэффициентов корреляции после удаления выбросов***

Рассмотрим увеличились ли показатели корреляции между переменными. У оценки туристического потока значение корреляции у всех коэффициентов уменьшилось. Оборот общественного питания принял такие значения корреляции, что три показателя снизилось и три увеличилось, а именно зависимость от оценки туристического потока, среднемесячной номинальной заработной платы и численности размещенных лиц в коллективных средствах размещения, стала меньше. Валовая добавленная стоимость туристической индустрии экономики приняла такие значения, что все показатели, кроме корреляции с числом турфирм, уменьшились. У среднемесячной номинальной заработной платы все показатели корреляции уменьшились, кроме значения, связанного с числом турфирм. У численности размещенных лиц в коллективных средствах размещения все значения корреляции упали. Валовой региональный продукт до удаления выбросов имел малые значения корреляции, сейчас некоторые из показателей увеличились, но связь так и осталась слабой. Число турфирм теперь имеет чуть большую связь со всеми показателями, кроме валового регионального продукта, здесь корреляция уменьшилась.

* 1. Матрица частных коэффициентов корреляции

Частные коэффициенты корреляции позволяют выявить «чистую» зависимость признака от одного из факторов и установить, каково было бы влияние этого фактора на величину признака при условии, что влияние других (другого) факторов на этот признак исключается. Если коэффициент равен 0, то взаимосвязь нейтральная и влияния нет, если значение находится в диапазоне от 0.09-0.19, то это говорит об незначительной связи. Значение в интервале 0.19-0.49 принимается, когда существует слабая связь, а в диапазоне 0.49-0.69 средняя взаимосвязь. Сильная связь имеется у показателей со значением частной корреляции равной 0.69-0.99

Матрица частных коэффициентов корреляции была построена с помощью Google Colaboratory, где используется язык питон. Матрица представлена на рисунке 2.6.1.



***Рисунок 2.5.1 –Матрица частных коэффициентов корреляции после удаления выбросов***

Видно, что в основном между показателями наблюдается средняя взаимосвязь, также есть и отрицательные показатели. Как показывает таблица оценка туристического потока имеет среднюю взаимосвязь с численностью размещенных лиц в коллективных средствах размещения. Оборот общественного питания и валовая добавленная стоимость туристической индустрии экономики имеют сильную взаимосвязь. Валовая добавленная стоимость туристической индустрии экономики имеют сильную взаимосвязь имеет среднюю связь с валовым региональным продуктом и числом турфирм. Среднемесячная номинальная заработная плата и валовой региональный продукт принимают такое значение, что можно говорить об средней связи между ними.

* 1. Сравнение частных и парных коэффициентов корреляции

Парный коэффициент корреляции показывает тесноту связи между двумя признаками на фоне действия остальных переменных, а частный характеризует взаимосвязь этих двух признаков при исключении влияния остальных переменных, т.е. их «личную» взаимосвязь.

Таким образом, если оказывается, что парный коэффициент корреляции между двумя переменными по модулю больше соответствующего частного, то остальные переменные усиливают связь между этими двумя признаками. Соответственно, если парный коэффициент корреляции между двумя переменными по абсолютной величине меньше частного, то остальные признаки ослабляют связь между рассматриваемыми двумя.

Если сравнивать значения, которые получились в представленных выше матрицах, то парные коэффициенты корреляции только в одной ячейке меньше частного коэффициента. Таким образом связь между числом турфирм и валовым региональным продуктом ослабевает из-за других факторов, остальные же значения, наоборот усиливаются.

* 1. Множественный коэффициент корреляции

Для измерения интенсивности совместного влияния всех факторов на изучаемый признак используют коэффициент множественной корреляции.

Множественных коэффициент корреляции был посчитан с помощью Excel, где были найдены определители через строенную функцию МОПРЕД() у матрицы парных коэффициентов после удаления выбросов. Использовалась формула , где - определитель всей матрицы, а -определитель матрицы коэффициентов корреляции между факторами.

Таким образом получилось значение равное 0,91. Если значение равно нулю, то это говорит о том, что все коэффициенты регрессионной модели кроме Y равны нулю. Из чего следует, что все регрессоры, включенные в модель не оказывают никакого влияния на отклик. Если множественный коэффициент получился равным 1, значит, отсутствуют случайные возмущения, а предсказанный отклик всегда будет совпадать с истинным откликов Y.

* 1. Вывод

После проведения корреляционного анализа было выявлено, что в основном все переменные усиливают связь между друг другом, кроме того, почти все показатели слабо зависят от валового регионального продукта. В исследовании была поставлена цель узнать от чего может зависеть оценка туристического потока, таким образом, можно сделать вывод, что данный показатель зависит от оборота общественного питания, валовой добавленной стоимость на туристическую индустрию экономики, численности размещенных лиц в коллективных средствах размещения и от числа турфирм. Неожиданным стало то, что нет зависимости от среднемесячной номинальной заработной платы, но вероятно это связано с тем, что нужно брать показатели реальной заработной платы, поскольку номинальная зарплата показывает какую сумму денег получает работник за месяц, а реальная заработная плата отвечает за количество товаров и услуг, которые можно приобрести за номинальную.

Интересным наблюдением в корреляционном анализе является этап сравнения между полученными результатами после и до удаления выбросов. Выше уже было сказано, что общая картина не изменилась, но в целом показатели корреляции в основном уменьшились.

# **Глава 3 Кластерный анализ**

В данной главе представлен кластерный анализ, который проводился на исходных данных, т.е на данных, имеющие выбросы. Это сделано из размышлений, что кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу, были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга, это значит, что все выбросы вероятно попадут в один кластер и не будут мешать исследованию.

1. 1. Разбиение на кластеры различными способами

Для разбиения имеющихся данных будут использованы следующие методы: межгрупповая связь, внутригрупповая связь, ближайший сосед, дальний сосед, центроидная кластеризация, медианная кластеризация и метод Варда. Рассмотрим, что из себя представляет каждый из методов.

Межгрупповая связь - дистанция между кластерами, которая равна среднему значению всех расстояний между всеми возможными парами точек из обоих кластеров. Информация, необходимая для расчёта дистанции, находится на основании всех теоретически возможных пар наблюдений.

Внутригрупповая связь - дистанция между двумя кластерами рассчитывается на основании всех возможных пар наблюдений, принадлежащих обоим кластерам, причём учитываются также и пары наблюдений, образующиеся внутри кластеров.

Ближайший сосед - дистанция между двумя кластерами определяется как расстояние между парой наблюдений, расположенных друг к другу ближе всего, причём каждое наблюдение берётся из своего кластера.

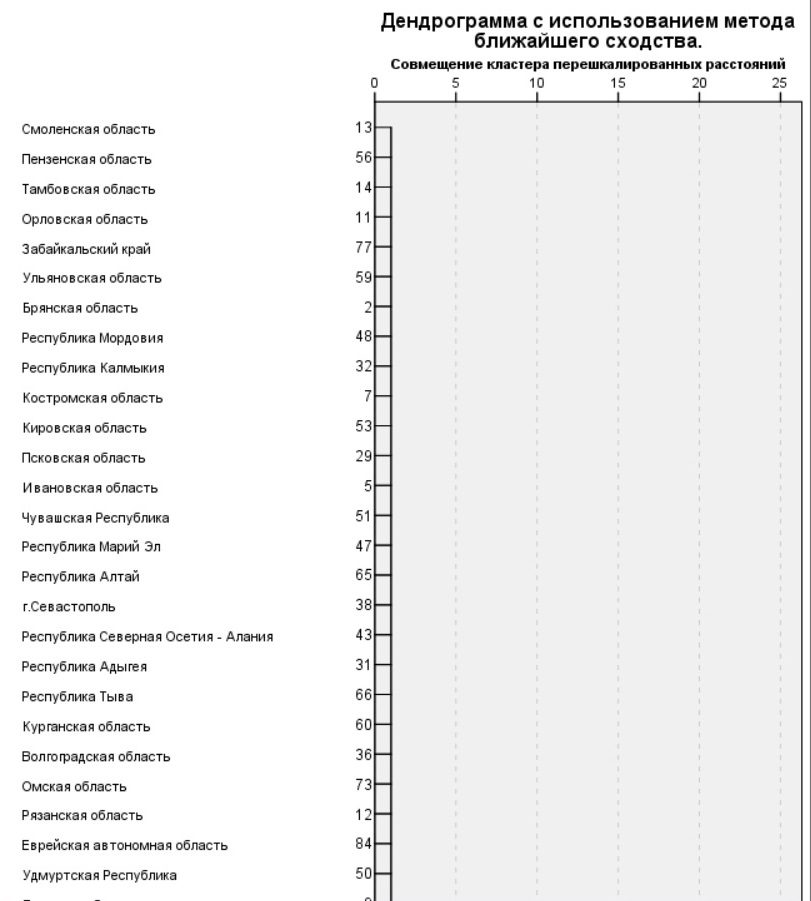
Самый дальний сосед - дистанция между двумя кластерами определяется как расстояние между самыми удалёнными друг от друга значениями наблюдений, причём каждое наблюдение берётся из своего кластера.

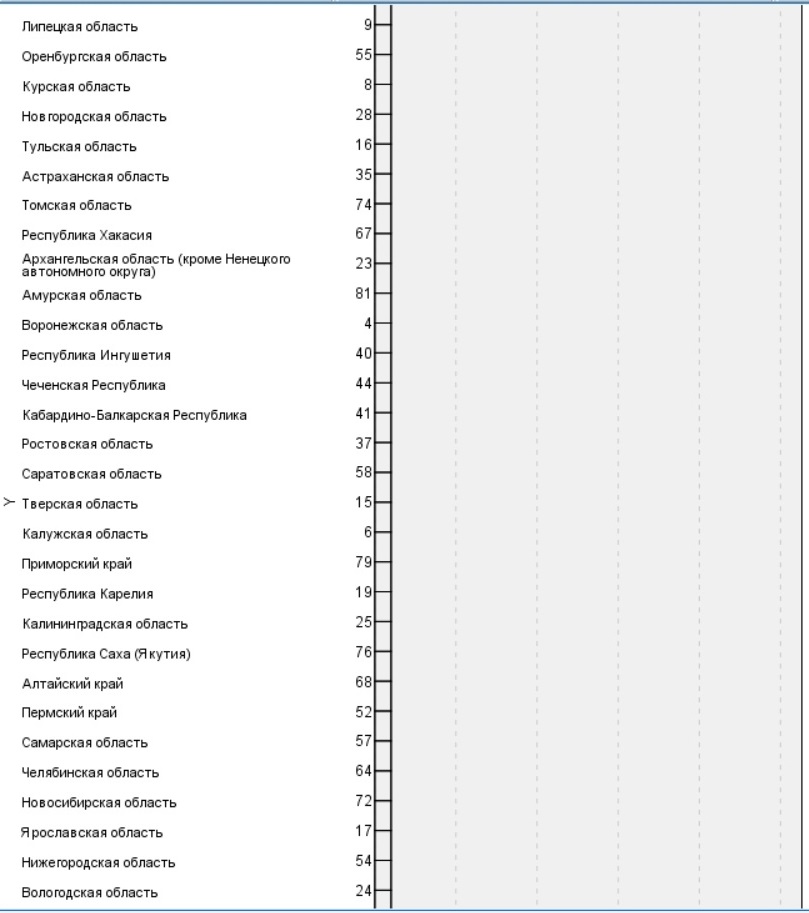
Центроидная кластеризация - в обоих кластерах рассчитываются средние значения переменных относящихся к ним наблюдений. Затем расстояние между двумя кластерами рассчитывается как дистанция между двумя осредненными наблюдениями.

Медианная кластеризация- тот же центроидный метод, но центр объединенного кластера вычисляется как среднее всех объектов.

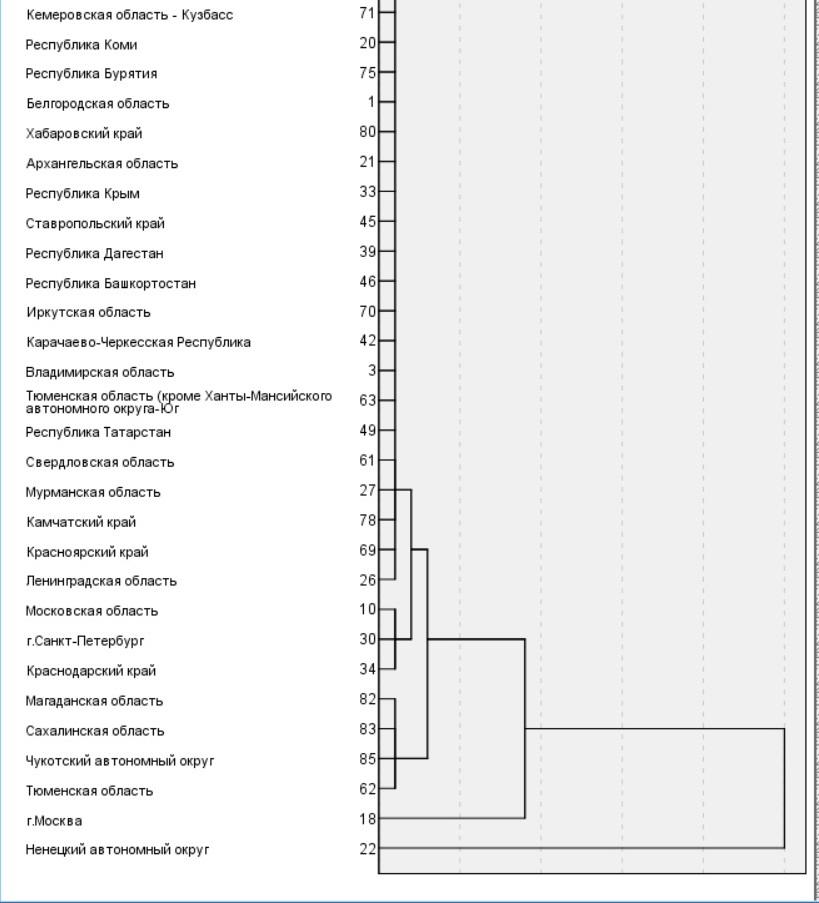
Метод Варда - сначала в обоих кластерах для всех имеющихся наблюдений производится расчёт средних значений отдельных переменных. Затем вычисляются квадраты евклидовых расстояний от отдельных наблюдений каждого кластера до этого кластерного среднего значения. Эти дистанции суммируются. Потом в один новый кластер объединяются те кластеры, которые дают наименьший прирост общей суммы дистанций.

Рассмотрим получившиеся дендрограммы. На дендрограмме номера объектов следуют по вертикали. По горизонтали отмечены расстояния (в условных единицах), на которых происходит объединение объектов в кластеры.

 ***Рисунок 3.1.1 –Первая часть дендрограммы при разбиении методом ближнего соседа***



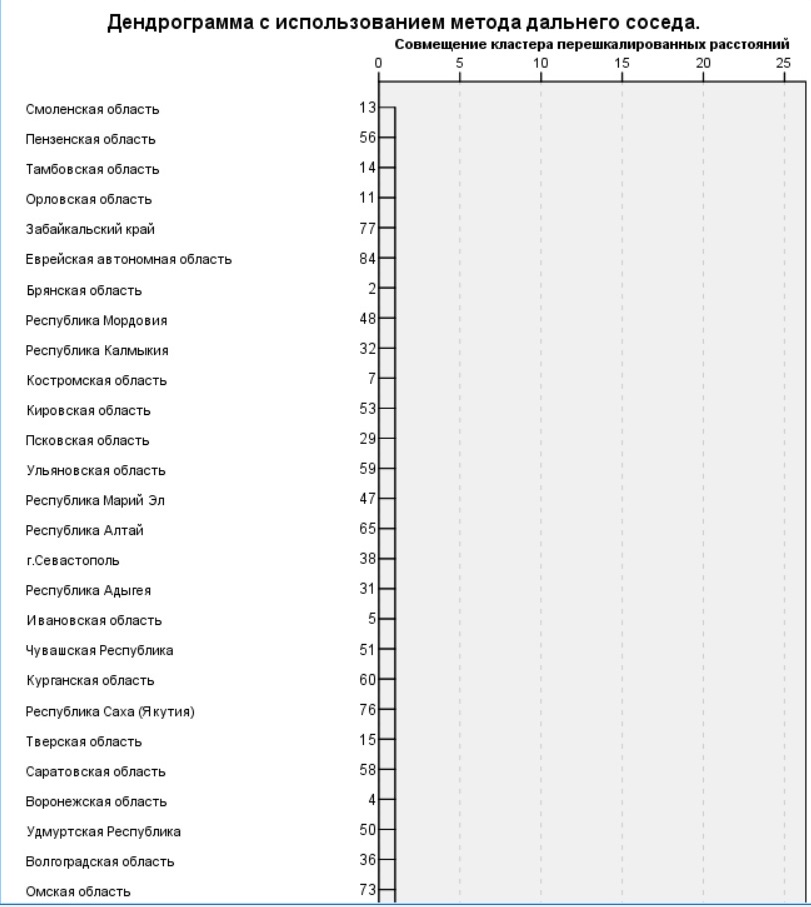
***Рисунок 3.1.2 –Вторая часть дендрограммы при разбиении методом ближнего соседа***



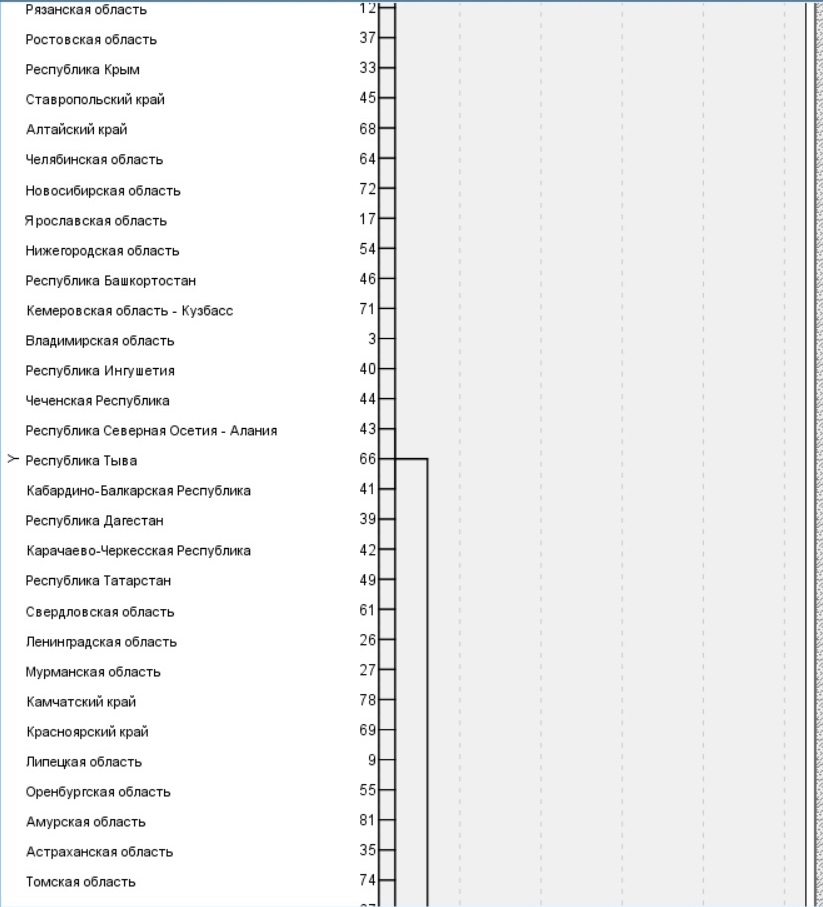
***Рисунок 3.1.3 –Третья часть дендрограммы при разбиении методом ближнего соседа***

На первых шагах происходит образование кластеров: первый кластер начиная со Смоленской области и заканчивая Ленинградской, второй кластер Московская область, г. Санкт-Петербург и краснодарский край, третий кластер Магаданская область, Сахалинская область, Чукотский автономный округ и Тюменская область. Далее образуется кластеры с помощью объединения первого и второго, на следующем шаге объеденные кластеры первый и второй объединились с третьим кластером. На последних шагах три объединённых кластера добирали объекты г. Москва и Ненецкого автономного округа, которые никуда не попали и таким образом получился один кластер. В целом, если рассматривать алгоритм формирования дендограммы, то нужно сказать, что алгоритм заканчивает действие, когда все объекты будут объединены в один кластер. В данном случае я бы выделила 3 кластера, но тогда 2 объекта мы потеряем.

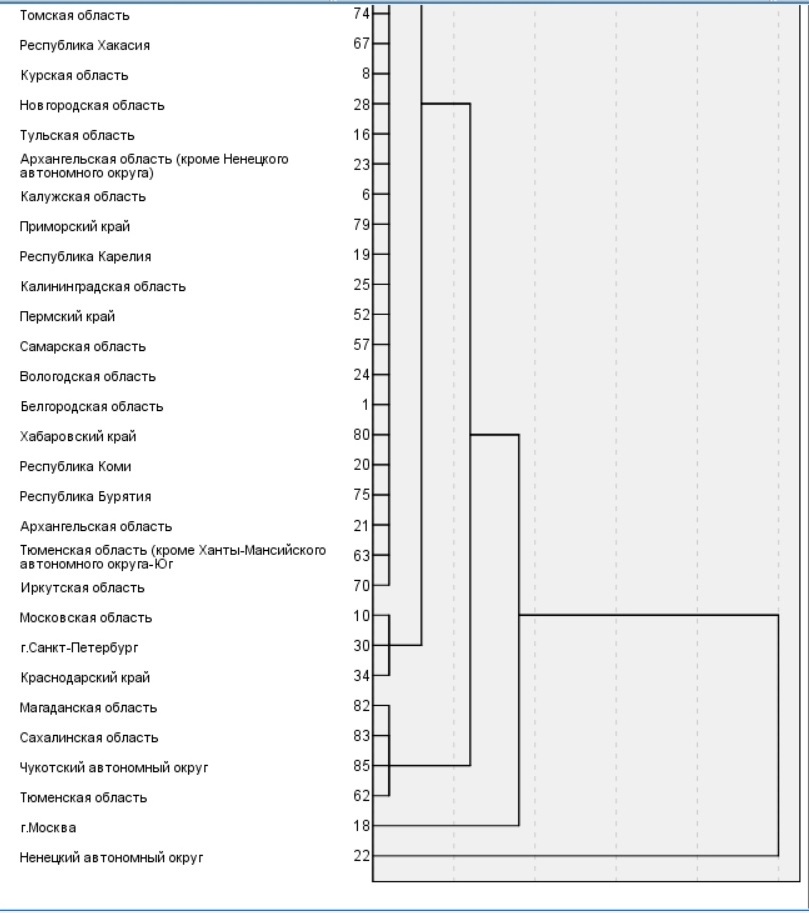
Перейдем к следующему методу, на рисунках 3.1.4,3.1.5,3.1.6 изображена дендрограмма для метода ближнего соседа.



***Рисунок 3.1.4 –Первая часть дендрограммы при разбиении методом дальнего соседа***



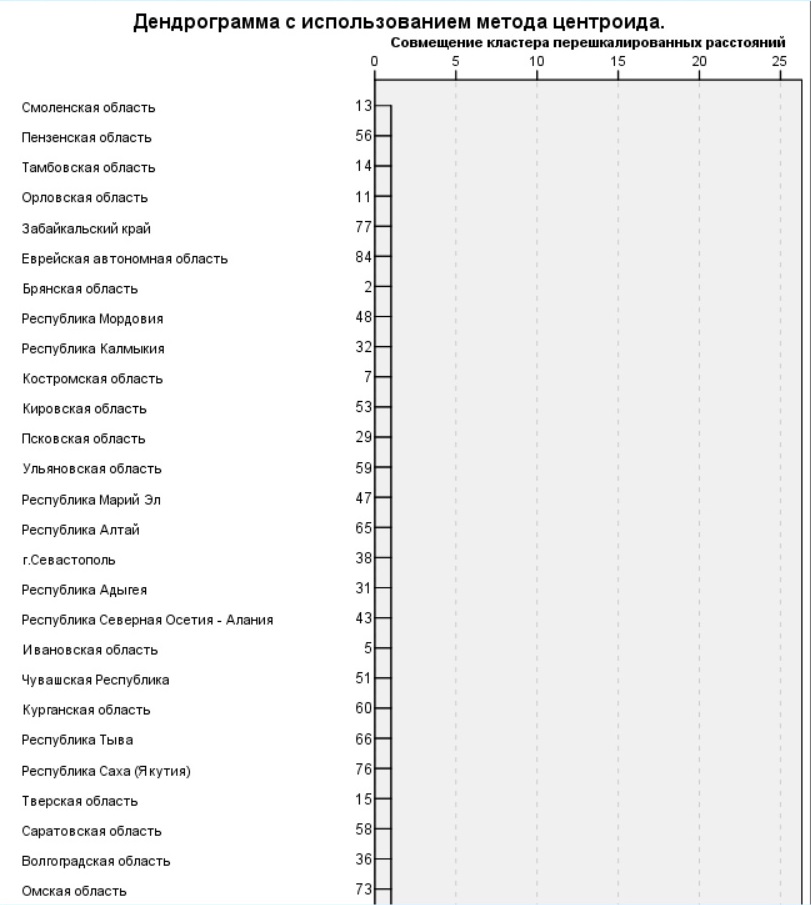
***Рисунок 3.1.5 –Вторая часть дендрограммы при разбиении методом дальнего соседа***



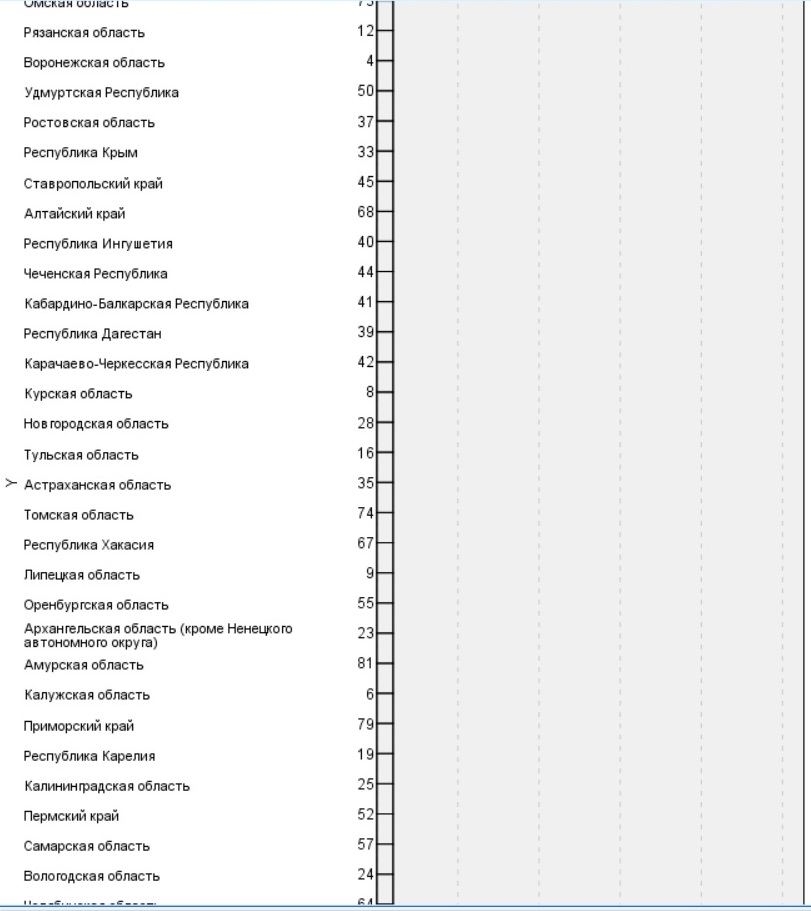
***Рисунок 3.1.6 –Третья часть дендрограммы при разбиении методом дальнего соседа***

Исходя их метода дальнего соседа кластеры формируются следующим образом, формируется 3 начальных кластера, а именно в первый входят объекты, начиная со Смоленской области и заканчивая Тюменской областью(не включающей в себя Ханты-Мансийский автономный округ), второй кластер включает в себя Московскую область, г. Санкт-Петербург и Краснодарский край, третий кластер состоит из Магаданскую область, Сахалинскую область и Чукотский автономный округ. На первом этапе кластеры поделились почти также как и при методе ближнего соседа, конкретно, второй и третий кластер совпадают в обоих случаях. Далее происходит аналогичное объединение первого и второго кластера через объект 66, позже к ним подключается третий кластер и оставшиеся два не сгруппированных объекта г. Москва и Ненецкий автономный округ. Исходя из вышесказанного можно выделить два кластера, а именно, объеденный первый со вторым и третий.

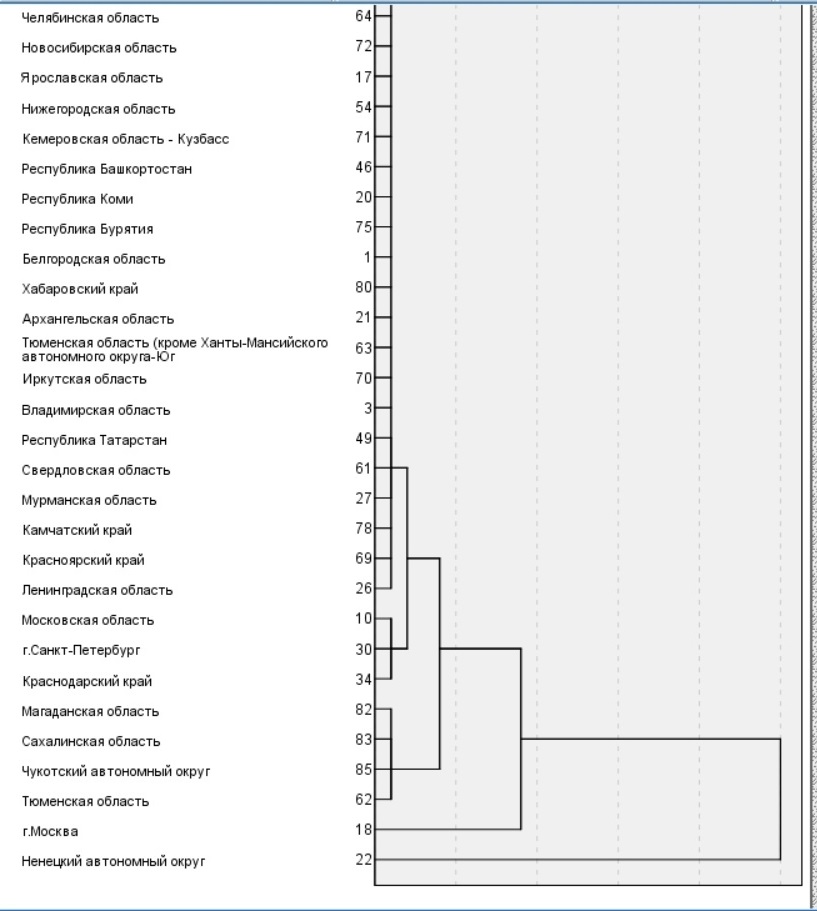
На рисунках 3.1.7,3.1.8,3.1.9 представлена дендрограмма, сгенерированная методом центроида.



***Рисунок 3.1.7 –Первая часть дендрограммы при разбиении методом центроида***



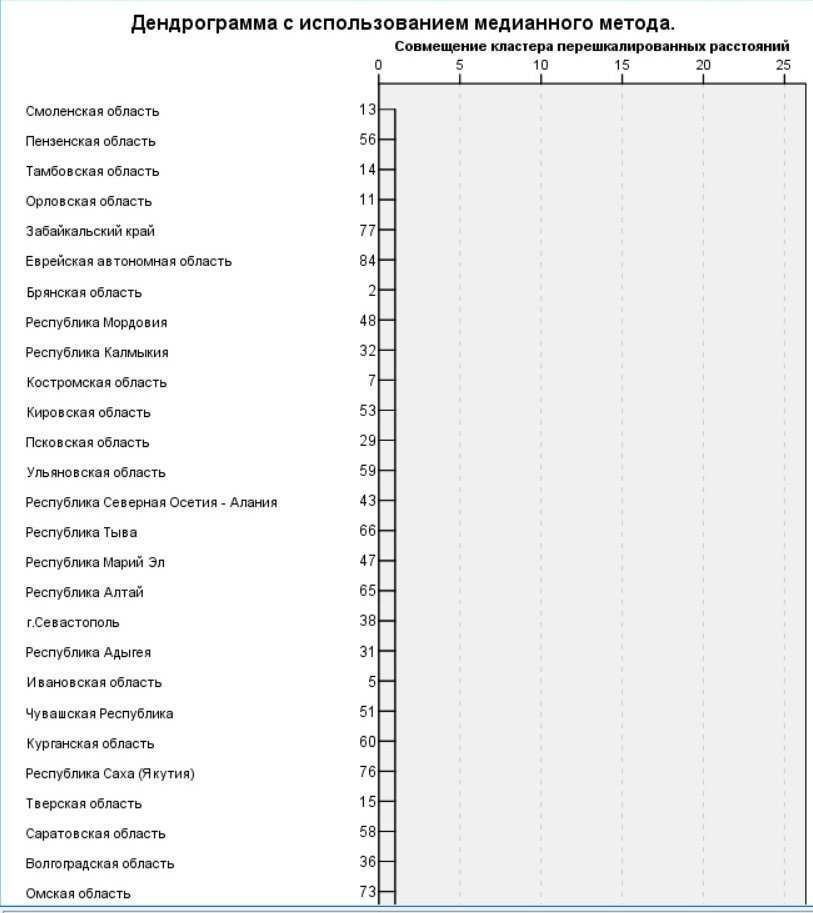
***Рисунок 3.1.8 –Вторая часть дендрограммы при разбиении методом центроида***



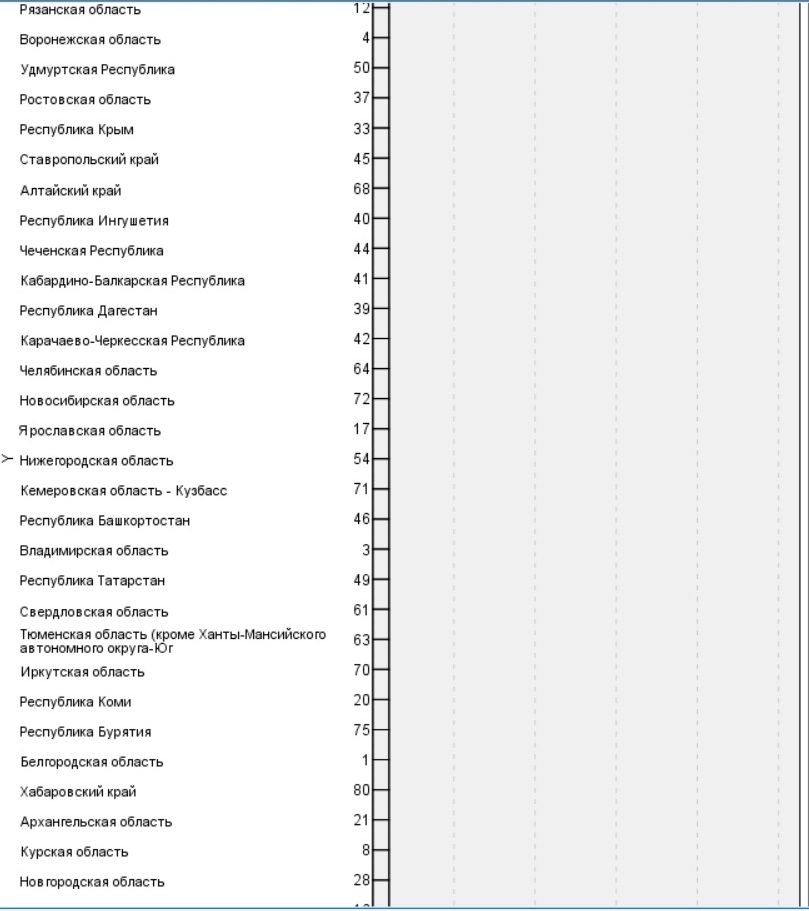
***Рисунок 3.1.9 –Третья часть дендрограммы при разбиении методом центроида***

Метод центроида получил идентичные результаты кластеризации, как и в случае с методом ближнего соседа, единственное отличие, что первый и второй кластер объединялись по значению 61. В данном случае можно также выделить 3 кластера, можно рассмотреть разбиение, где берутся объединенный второй и первый, а также третий.

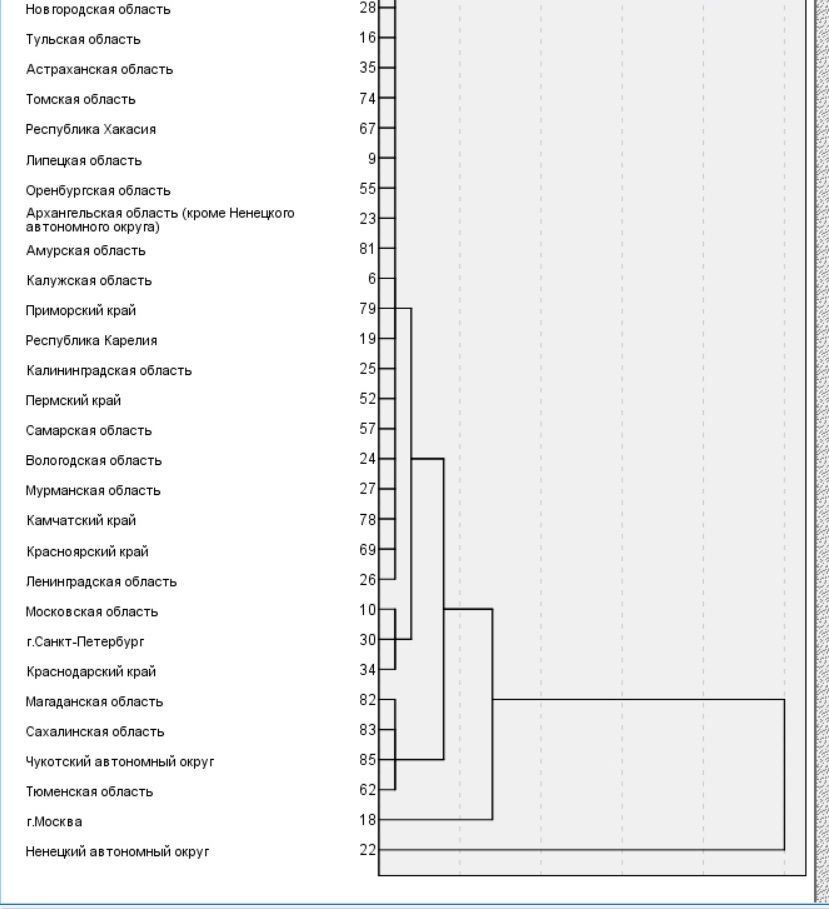
На рисунках 3.1.10,3.1.11,3.1.12 изображена дендрограмма, сделанная по медианному методу



***Рисунок 3.1.10 –Первая часть дендрограммы при разбиении медианным методом***

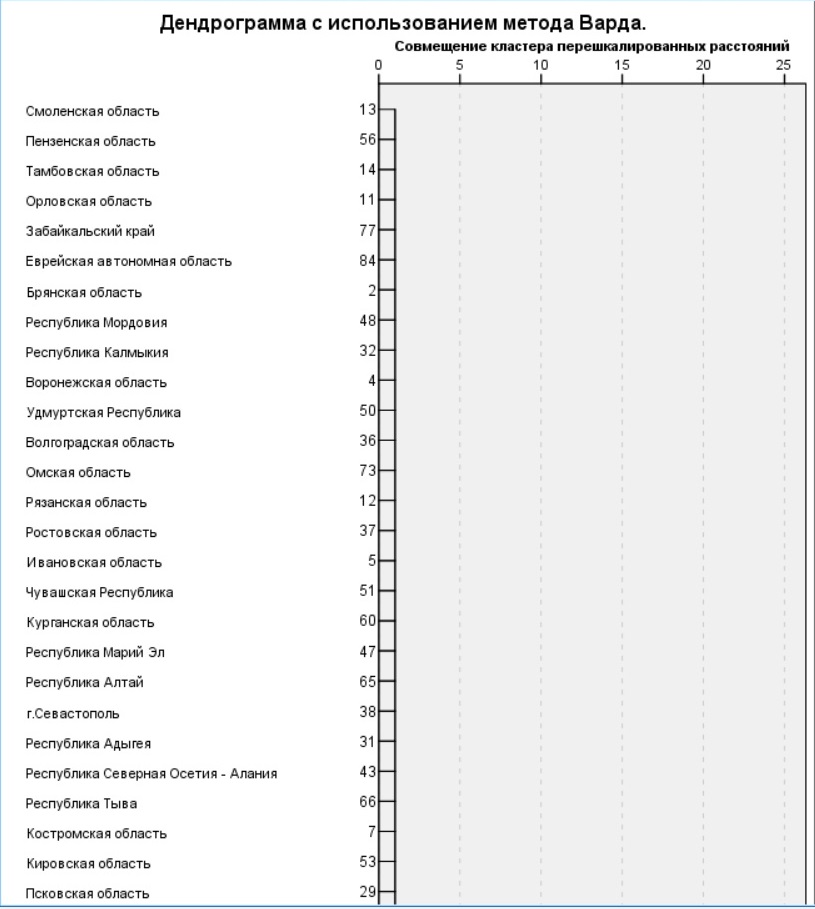


***Рисунок 3.1.11 –Вторая часть дендрограммы при разбиении медианным методом***

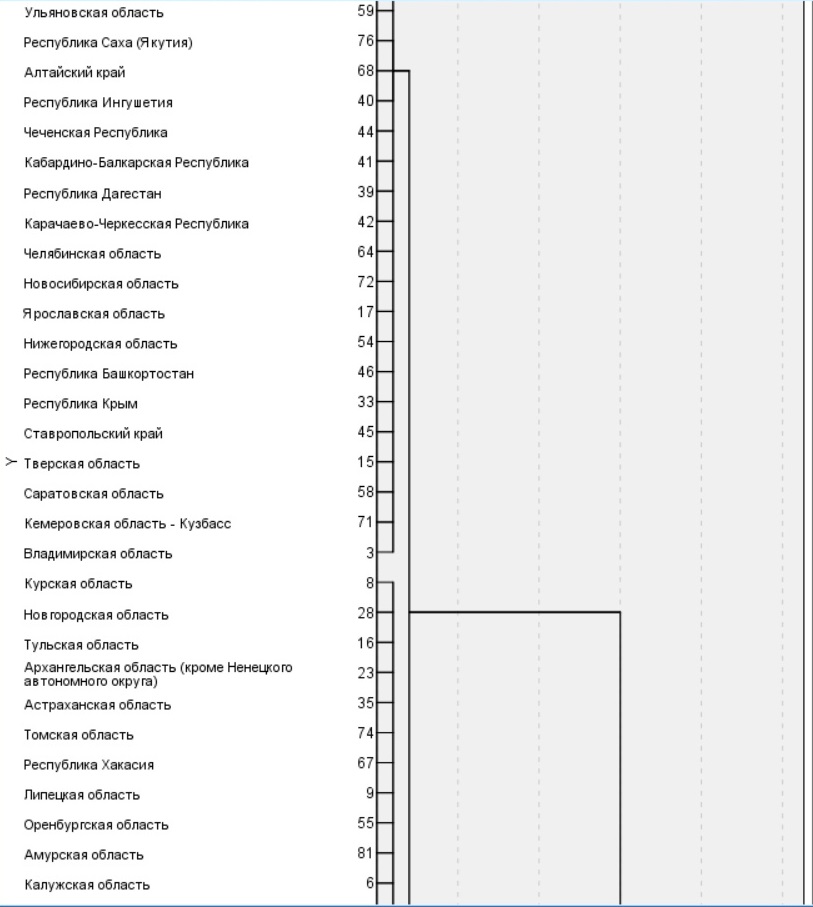
 ***Рисунок 3.1.12 –Третья часть дендрограммы при разбиении медианным методом***

Кластеризация медианным методом не сильно отличается от предыдущих, в данном случае можно аналогично выделить либо 2, либо 3 кластера.

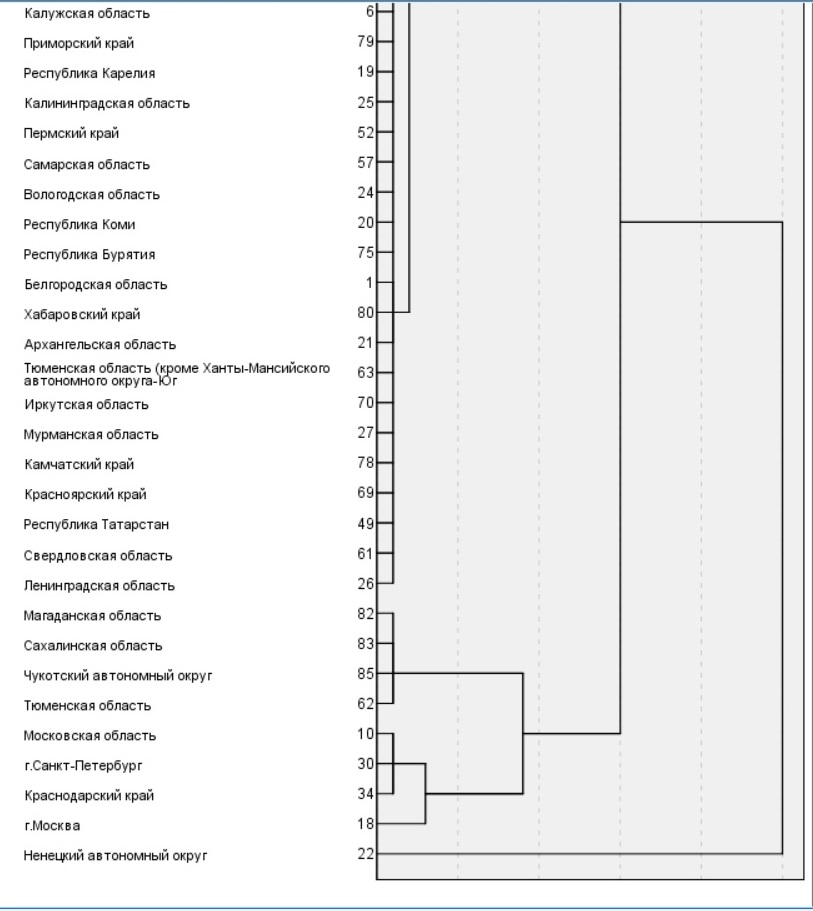
Рассмотрим результаты метода Варда на рисунках 3.1.13,3.1.14,3.1.15.



***Рисунок 3.1.13 –Первая часть дендрограммы при разбиении методом Варда***

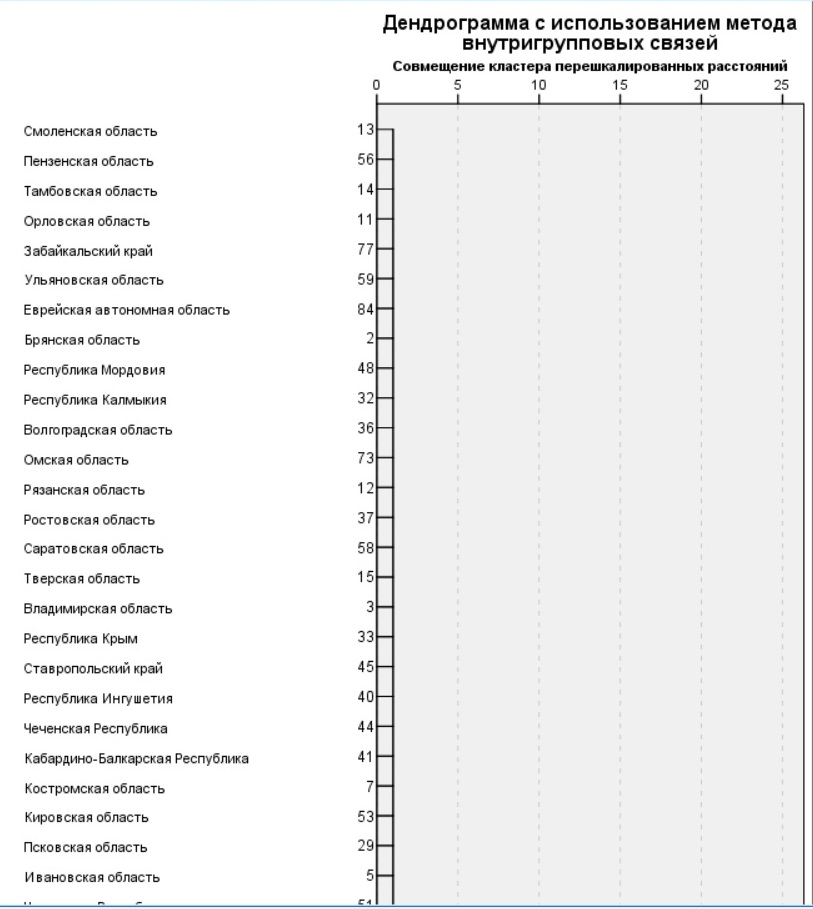


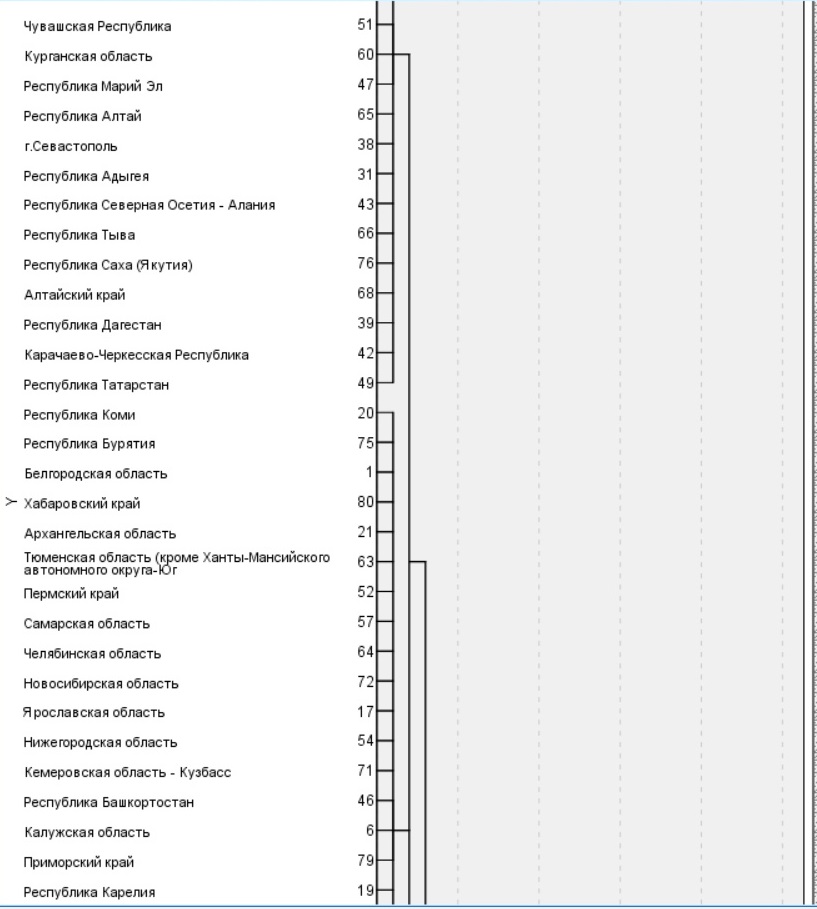
***Рисунок 3.1.14 –Вторая часть дендрограммы при разбиении методом Варда***

 ***Рисунок 3.1.15 –Третья часть дендрограммы при разбиении методом Варда***

На первом этапе образовалось 4 кластера, первый кластер включает в себя объекты начиная со Смоленской области и заканчивая Владимирской областью, во второй кластер вошли объекты от Курской до Ленинградской области, третий кластер — это Магаданская область, Сахалинская область, Чукотский автономный округ и Тюменская область, четвёртый кластер Московская область, г. Санкт-Петербург и Краснодарский край. На следующем этапе кластеры начали объединяться, так объединились первый и второй кластер, далее четвертый кластер с объектами г. Москва. На третьем этапе мы наблюдаем, что объединенный кластер четыре и г. Москва объединились с третьим кластером. На последних двух этапах происходит формирование одного единого кластера. В данном случае выгоднее остановиться на трех кластерах объединённый четвёртый с г. Москва и объеденные первый и второй и отдельный третий кластер.

Перейдем к результатам метода внутригрупповой связи, который представлен на рисунках 3.1.16,3.1.17,3.1.18.

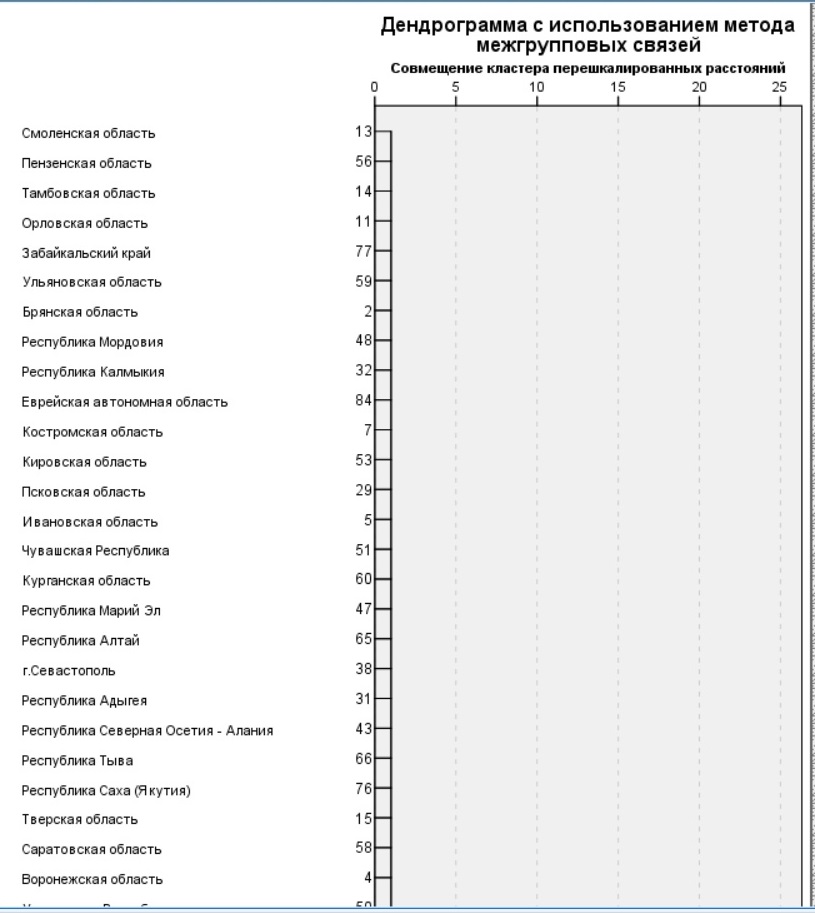
 ***Рисунок 3.1.16 –Первая часть дендрограммы при разбиении методом внутригрупповой связи***

 ***Рисунок 3.1.17 –Вторая часть дендрограммы при разбиении методом внутригрупповой связи***

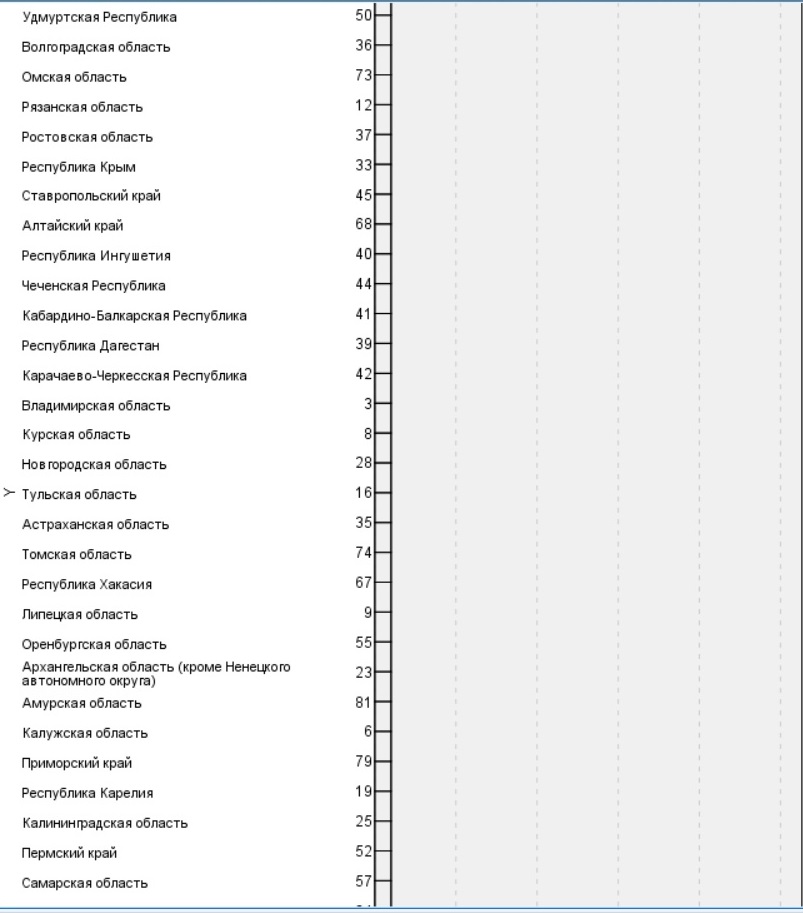
 ***Рисунок 3.1.18 –Третья часть дендрограммы при разбиении методом внутригрупповой связи***

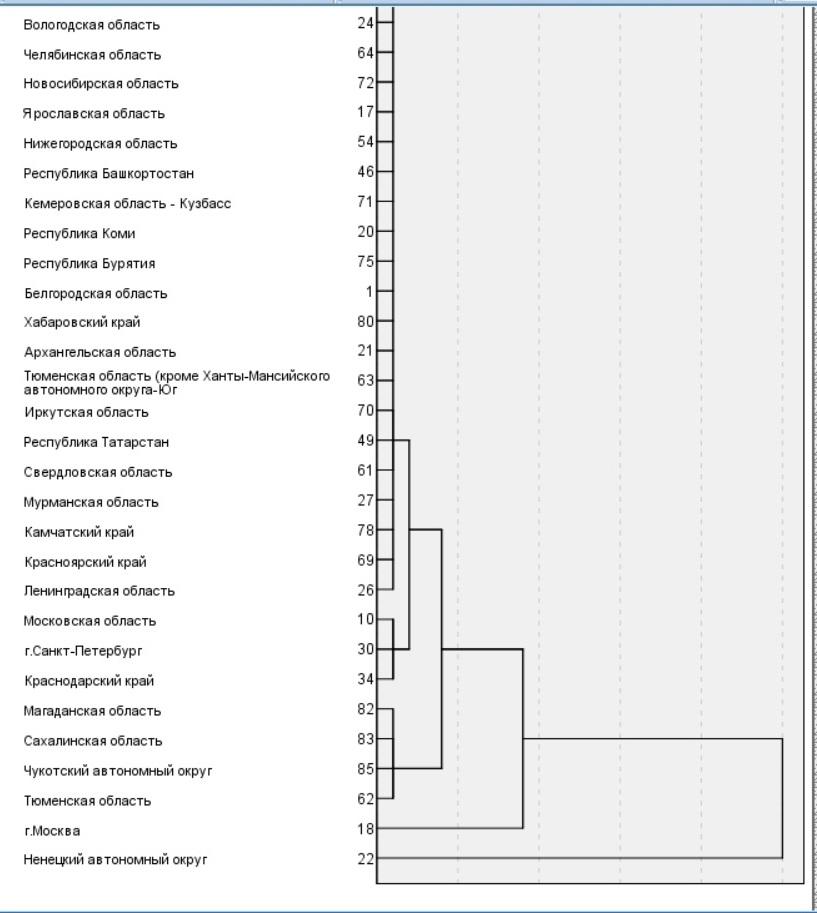
С помощью метода внутригрупповой связи получился интересный результат кластеризации. На первом этапе сформировалось 4 кластера, далее начало происходить объединение и формирования новых кластеров. Скорее всего в этом случае лучше остановиться на объединении кластеров на 5 этапе и рассматривать их.

Рассмотрим метод межгрупповой связи, который представлен на рисунках 3.1.19,3.1.20,3.1.21.



***Рисунок 3.1.19 –Первая часть дендрограммы при разбиении методом межгрупповой связи***

 ***Рисунок 3.1.20 –Вторая часть дендрограммы при разбиении методом межгрупповой связи***

 ***Рисунок 3.1.21 –Третья часть дендрограммы при разбиении методом межгрупповой связи***

Результаты кластеризации методом межгрупповой связи аналогичны тем, что были получены при методе ближнего соседа.

* 1. Выбор предпочтительно числа кластеров

Для того чтобы определить предпочтительное число кластеров в иерархической кластеризации в качестве критерия используются расстояния. Необходимо смотреть на коэффициент в протоколе объединения (расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учётом предусмотренного преобразования значений). Когда мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить. Иначе будут объединены кластеры, находящиеся на большом расстоянии друг от друга.

Таким образом, было определено, что предпочтительное число кластеров 2-3.

Вывод

По итогам исследовательской работы были достигнуты все цели и выполнены все задачи. Были продемонстрированы различные графические представления данных, рассмотрены основные характеристик (характеристики положения, характеристики разброса и ранговые характеристики), из предварительного анализа были найдены выбросы (как графически, так и теоретически через правило 1.5IQR И 3 сигм), а также стало ясно, что почти у всех показателей неоднородные данные, которые распределены неравномерно и имеют несколько мод. Кроме того, был проведен корреляционный анализ, в котором были рассмотрены зависимости между переменными до и после удаления выбросов. По итогу корреляционного анализа стало ясно, что туристический поток зависит от оборота общественного питания, валовой добавленной стоимость на туристическую индустрию экономики, численности размещенных лиц в коллективных средствах размещения и от числа турфирм. Кластерный анализ предоставил возможность рассмотреть большое количество методов для кластеризации данный, несколько методов дали одинаковый результат, но также были и те, кто заинтересовал, например, метод внутригрупповой связи, который вероятно лучше использовать в качестве наилучшего варианта. Итогом кластерного анализа является разбиение всех данных на 3 кластера. Таким образом все имеющиеся области и отдельные города по данным

оценки туристического потока, оборота общественного питания на душу, валовой добавленной стоимости туристической индустрии, среднемесячной номинальной заработной платы по полному кругу организация, численности размещенных лиц в коллективных средствах размещения, валового региональный продукт на душу населения и числа турфирм можно объединить по их значениям и получить более сжатый массив данных.

Литература и используемые информационные источники

Данные для исследования собирались на сайте Росстат:

<https://rosstat.gov.ru/folder/10705>

Приложение А

**Все операции в SPSS.**