**Федеральное агентство по образованию Российской Федерации**

**Государственное образовательное учреждение**

**высшего профессионального образования**

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**Кафедра прикладной математики**

Преподаватель,

д.т.н. А.А. Халафян

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ОТЧЕТ**

**СТАТИСТИЧЕСКИЙ АНАЛИЗ НЕКОТОРЫХ ХАРАКТЕРИСТИК МИРОВЫХ ГОСУДАРСТВ.**

Работу выполнил студент 4 курса факультета компьютерных технологий и прикладной математики Демьяненко Н.П.  
спец. 010304 – Прикладная математика и информатика

Краснодар 2023

СОДЕРЖАНИЕ

[**1 Исходные данные 3**](#_Toc153748988)

[**2 Графический анализ 5**](#_Toc153748989)

[**2.1 2D Graphs 5**](#_Toc153748990)

[**2.1.1 2D Histogramms 5**](#_Toc153748991)

[**2.1.2 2D Scatterplots 7**](#_Toc153748992)

[**2.1.3 2D Box Plots 12**](#_Toc153748993)

[**2.2 3D SequentialGraphs 15**](#_Toc153748994)

[**2.2.1 Box Plots 15**](#_Toc153748995)

[**3 Основные статистики 16**](#_Toc153748996)

[**3.1 Описательные статистики 16**](#_Toc153748997)

[**3.2 Корреляционная матрица 16**](#_Toc153748998)

[**3.3 Критерий Стьюдента сравнения средних 16**](#_Toc153748999)

[**3.3.1 t-test, independent, by groups 17**](#_Toc153749000)

[**3.4 Группировка и однофакторная ANOVA 17**](#_Toc153749003)

[**4 Частотный анализ 21**](#_Toc153749004)

[**4.1 Таблицы частот 21**](#_Toc153749005)

[**4.1 Таблицы кросстабуляции 25**](#_Toc153749005)

**5 Корреляционный анализ**  [**29**](#_Toc153749004)

[**6 Дисперсионный анализ 33**](#_Toc153749004)

[**7 Линейное многомерное моделирование взаимосвязей 35**](#_Toc153749004)

[**8 Нелинейное многомерное моделирование взаимосвязей 37**](#_Toc153749006)

[**8.1 Fixed Nonlinear Regression 37**](#_Toc153749007)

[**8.2 Пробит регрессия 38**](#_Toc153749008)

[**8.3 Экспоненциальная регрессия 38**](#_Toc153749009)

[**9 Дискриминантный анализ 42**](#_Toc153749011)

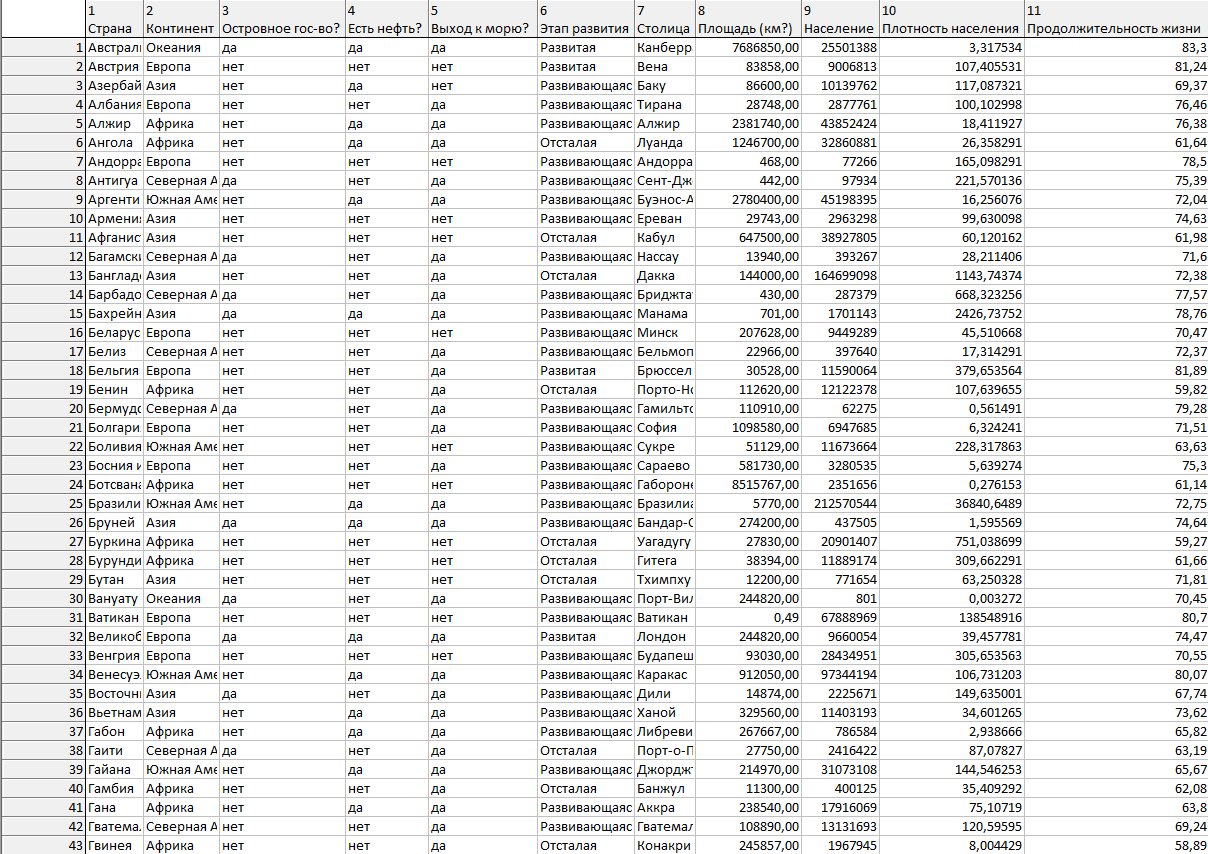
[**10 Кластерный анализ 50**](#_Toc153749011)

[**10.1 Метод к-средних 50**](#_Toc153749011)

[**10.2 Древовидная кластеризация 54**](#_Toc153749011)

[**11Факторный анализ 58**](#_Toc153749011)

**1 Исходные данные**

В исходной таблице представлены данные по старанам мира.

Показатели:

1. Страна – Название государства
2. Континент – Континент на котором располагается государство
3. Островное гос-во? – Является ли государство островным
4. Есть нефть? – Производится ли нефтедобыча в государстве
5. Выход к морю? – Имеет ли государство выход к мировому океану
6. Этап развития – Развитость государства по версии ООН
7. Столица – Столица государства
8. Площадь (км²) – Площадь контролируемой суши государством
9. Население – Количество людей проживающих в государстве
10. Плотность населения - Население/Площадь (км²)
11. Продолжительность жизни – Средняя продолжительность жизни человека в государстве

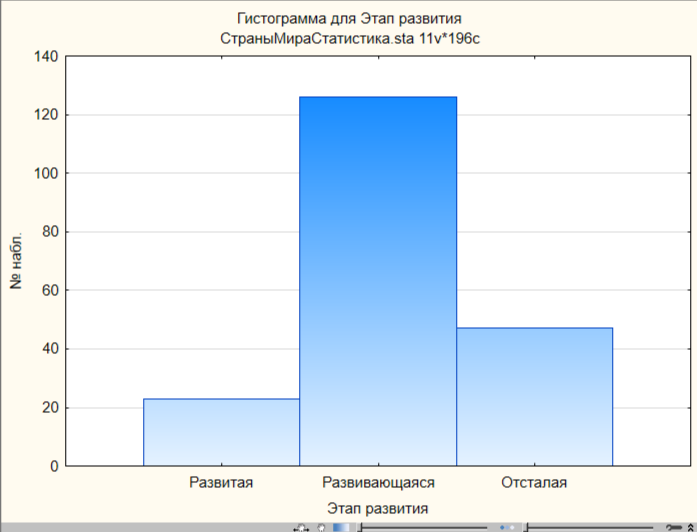
**2 Графический анализ**

**2.1 2D Graphs**

**2.1.1 2D Histogramms**

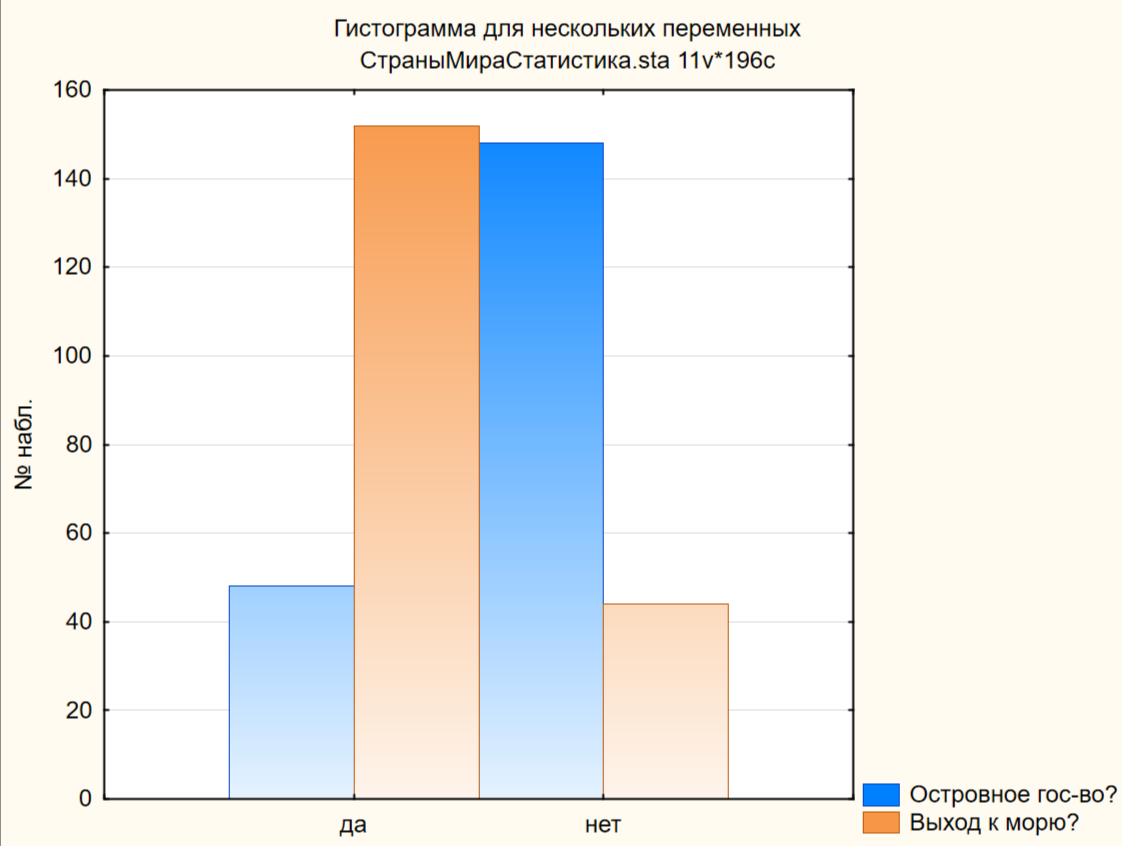
2D Histogramms являются графическими представлениями распределения частот выбранных переменных.

2D Histogramms Regular (простые) – столбчатая диаграмма распределения частот.



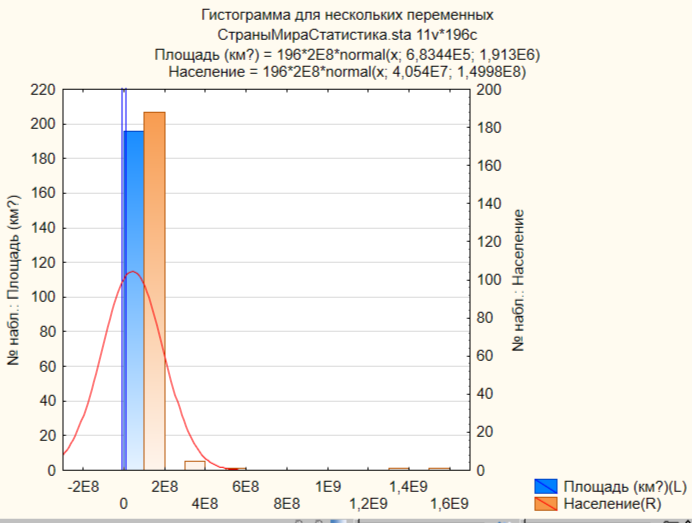
Данная 2D Histogramms Regular построена по этапам развития государств. Из гистограммы видно, что больше всего встречаются развивающиеся страны, более 120.

2D Histogramms Multiple (составные) – изображают распределение частот для нескольких переменных на одном графике.



Данная 2D Histogramms Multiple построена по характеристикам: островное ли государство и имеет ли оно выход к морю. Из гистограммы видно, что государств расположенных на материке и имеющих выход к морю значительно больше островных государств, а количество не имеющих выхода – сопоставимо с числом островных государств.

2D Histogramms Double-Y (с двойной осью Y) – Комбинация двух по-разному масштабированных составных гистограмм.



Данная 2D Histogramms Double-Y построена по двум переменным: площадь и население. Из гистограммы видно, что чаще всего встречаются государства площадью и населением в интервале от 0 до 200 млн.

**2.1.2 2D Scatterplots**

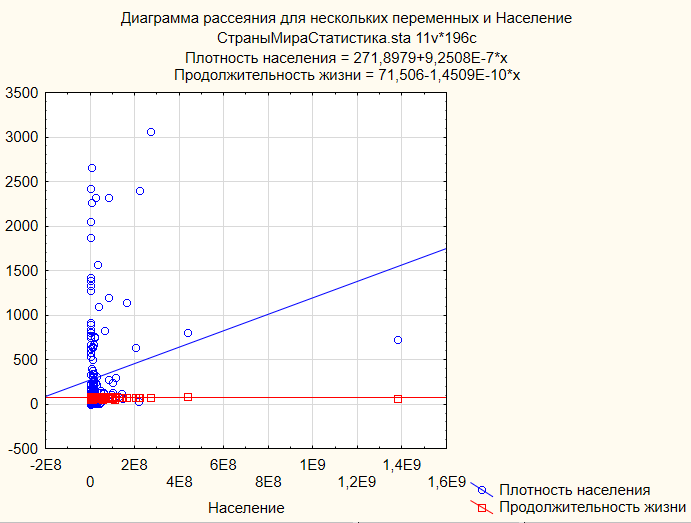
2D Scatterplots (диаграммы рассеяния) визуализируют зависимость между двумя переменными.

2D Scatterplots Regular (простые) - визуализируют зависимость между двумя переменными X и Y.



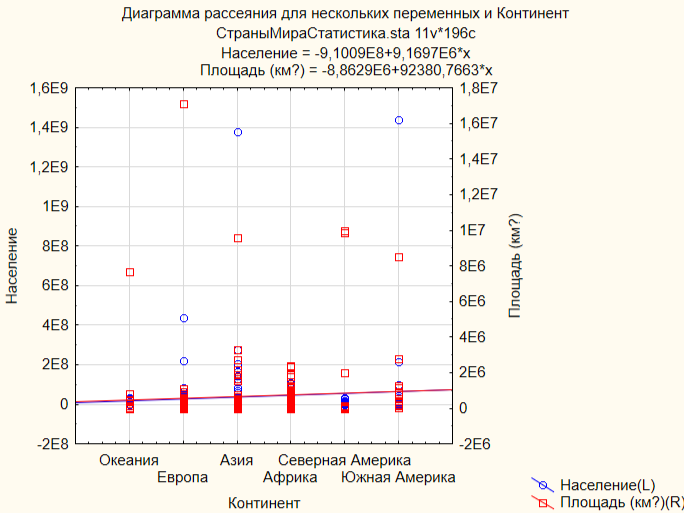
Данная 2D Scatterplots Regular диаграмма отображает зависимости между продолжительностью жизни и плотностью населения. В среднем, продолжительность жизни выше в странах с более высокой плотностью населения.

2D Scatterplots Multiple (составные) – состоит из нескольких зависимостей и изображает несколько корреляций.



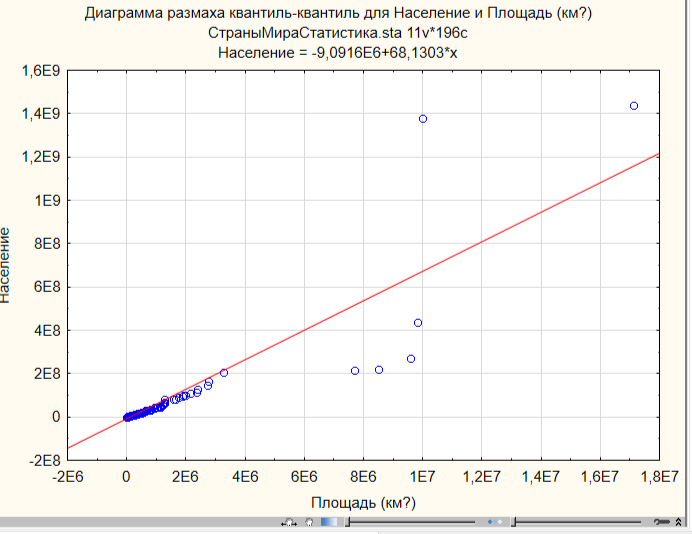
Данная 2D Scatterplots Multiple диаграмма отображает зависимости между плотностью населения, населением и продолжительностью жизни. Тут сложно выявить какую либо зависимость из-за слабо сопоставимых данных.

2D Scatterplots Double-Y (с двойной осью Y) – комбинация двух составных диаграмм рассеяния для одной переменной X и двух различных наборов переменных Y.



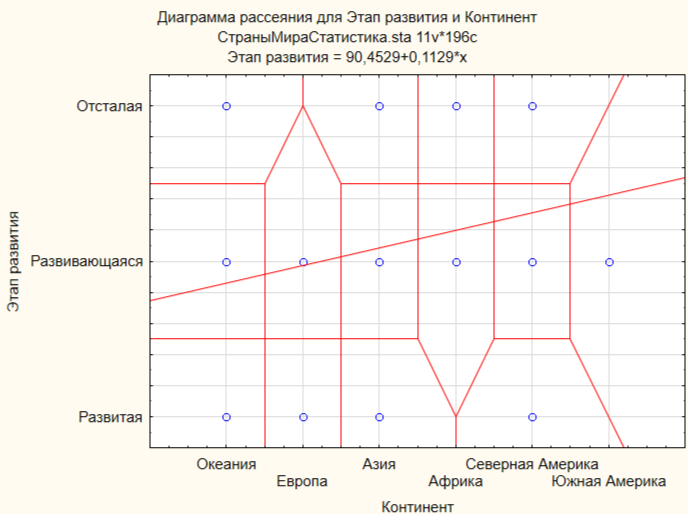
Данная 2D Scatterplots Double-Y диаграмма отображает комбинацию двух составных диаграмм: первая зависимость между населением и континент, вторая зависимость площадь и континентом.

2D Scatterplots Quartile (квантилей) – показывает зависимость между квантилями двух переменных, позволяющая оценить сходство эмпирических распределений.



Данная 2D Scatterplots Quartile отображает зависимость между населением и площадью. Можно сделать вывод, что корреляция между данными переменными является средней.

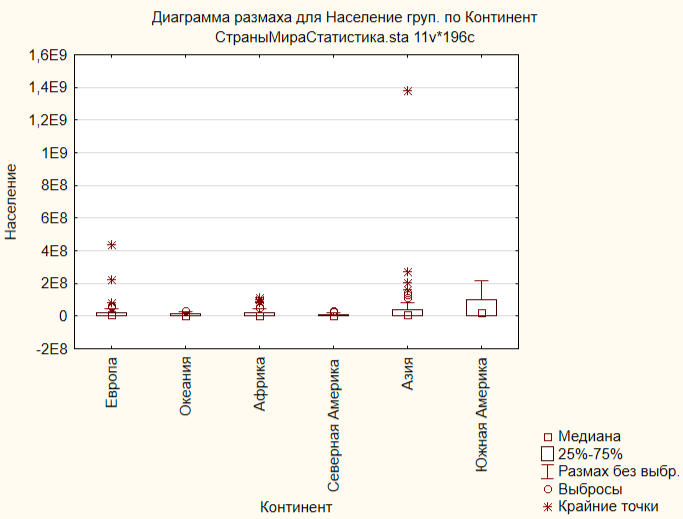
2D Scatterplots Voronoi (Вороного) – показывает разделение пространства между точками данных. Пространство между отдельными точками делится границами на такие области, каждая точка которых находится ближе к заключенной внутри точке, чем к любой другой соседней точке.



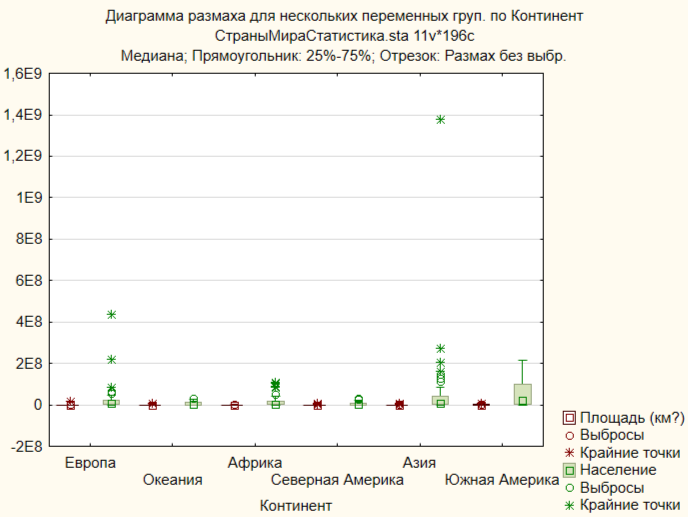
Данная 2D Scatterplots Voronoi диаграмма построена для переменных этапа развития и континента.

**2.1.3 2D Box Plots**

2D Box Plots (графики ящика – диаграммы размаха) – на этих диаграммах изображаются диапазоны или характеристики распределения значений выбранной переменной отдельно по группам, заданным категориальной переменной.



На этой 2D Box Plots Regular диаграмме представлены характеристики распределения значений переменной населения по группам континентам Можно сделать выводы, что медианное население в Южной Америке в два раза больше остальных континентов. Так же можно отметить, что самая большая крайняя точка находится в Азии.



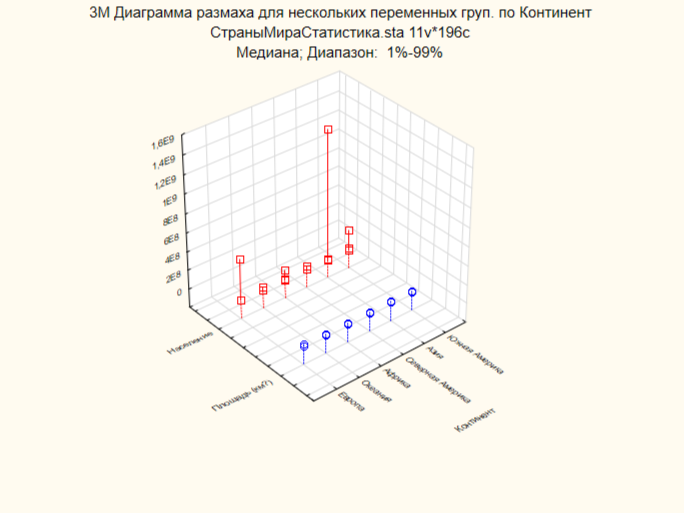
На этой 2D Box Plots Multiple диаграмме представлены характеристики распределения значений переменных населения и площадипо группам континентов. Из-за не соразмерно меньшей площади, в сравнении с населением, данная диаграмма не показательнее предыдущей.

**2.2 3D SequentialGraphs**

**2.2.1 Box Plots**

Box Plots - это диаграммы размаха.

Построим диаграмму размаха для переменных население и площадь по группам континент.

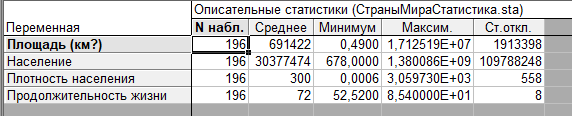


По данной диаграмме можно сделать вывод, что в Европе и Азии есть представители стран с крупным населением.

3 **Основные статистики**

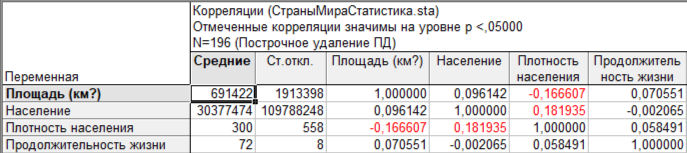
**3.1 Описательные статистики**

Таблица описательных статистик для стран:



Здесь показаны средние, минимальные и максимальные значения количественных переменных, а так же средние отклонения.

**3.2 Корреляционная матрица**



Из таблицы видно, все значения. Кроме диагональных, меньше 0,25 и не отмеченные красным - близки к нулю, что говорит о слабой корреляции, данные слабо зависят друг от друга.

**3.3 Критерий Стьюдента сравнения средних**

**t-test, independent**, by groups применяется если сравниваются средние количественной переменной значения которой разбиваются группирующей переменной на 2 группы.

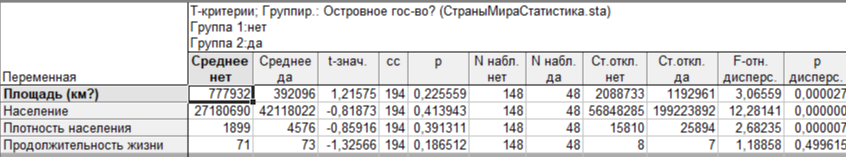
**t-test, independent**, by variables применяется, если сравниваются 2 столбца одной и той же величины, но записанные для разных объектов.

**t-test, dependent** samples применяем, если сравнивается одна и та же величина при 2 повторных измерениях.

**t-test, single samples** применяем, если сравниваем среднее группы с средним генеральной совокупности.

**3.3 t-test, independent, by groups**

Проверим равенство средних для площади, населения, его плотности и продолжительности жизни по группам является ли государство островным.

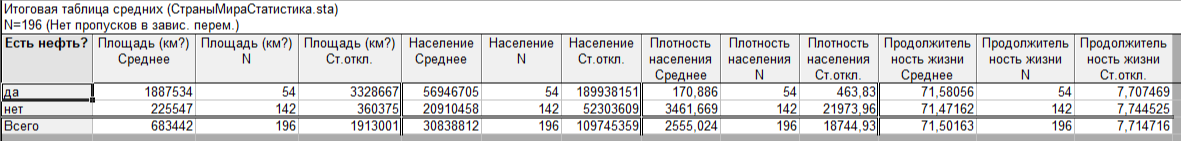


По данным таблицы можно сделать вывод, что данные статистически не значимы. Т.к. p < 0.05. гипотеза о равенстве дисперсии не выполняется для всех переменных, кроме продолжительности жизни.

**3.4 Группировка и однофакторная ANOVA**

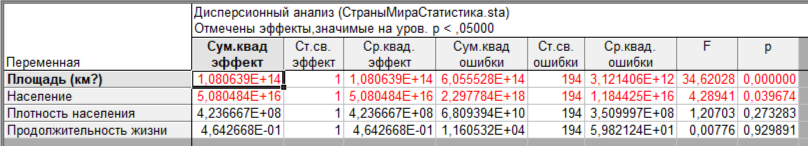
Модуль Breakdown & one-way ANOVA (группировка и однофакторный дисперсионный анализ *ANOVA* определяет внутригрупповые описательные статистики и корреляции для зависимых переменных в каждой из нескольких групп, определенных группирующей переменной. Сравнивает средние и определяет, в каких именно группах отличие средних статистически значимо отличаются между собой. В качестве нулевой гипотезы предполагается, что средние в генеральной совокупности равны. В терминологии пакета группирующую переменную называют фактором.

Сравним средние для переменных площади, населения, его плотности и продолжительности жизни по группам государств в которых есть нефть.

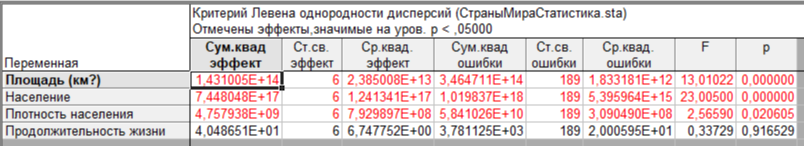
****

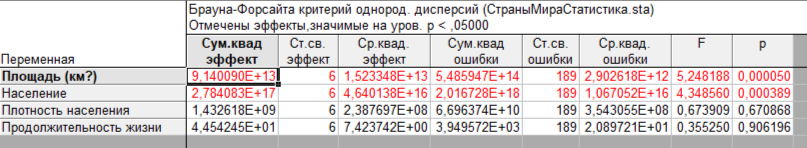
По данной таблице можно сделать следующие выводы:

В среднем площадь и население стран с нефтью больше, чем без неё, при этом плотность населения ниже, а продолжительность жизни статистически не отличаются.



р критерия Фишера Дисперсионного анализа для площади и населения <0,05, то это означает, что могут быть группы объектов, для которых не верна гипотеза о равенстве средних.

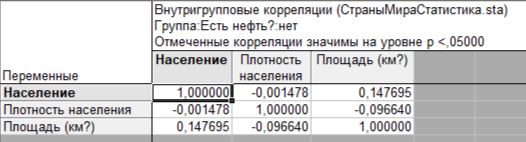




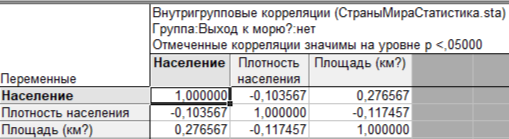
Критерии Левена и Брауна-Форсайта проверяют гипотезу о равенстве (однородности) дисперсий в 3 группах. Нулевая гипотеза: дисперсии в группах равны. Для переменных площадь и население (в Критерии Левена – плотность населения) p<0.05, это означает что нулевая гипотеза не верна.

Критерий Левена и критерий Брауна-Форсайта являются статистическими тестами, используемыми для проверки гомогенности дисперсии в выборке. Однако, главным отличием между ними является то, что критерий Левена используется для небольших выборок (до 200 элементов), а критерий Брауна-Форсайта - для больших выборок (более 200 элементов). Критерий Левена также имеет две разновидности: для нормального распределения и для ненормального распределения. Он основан на сравнении медиан двух групп, полученных путем разделения выборки на две части. Критерий Брауна-Форсайта основан на рангах элементов выборки и позволяет проверить гипотезу о равенстве дисперсии в нескольких группах. Он имеет более высокую мощность, чем критерий Левена, и может использоваться даже при наличии значительных отклонений от нормального распределения. Таким образом, выбор между критерием Левена и критерием Брауна-Форсайта зависит от размера выборки и предполагаемого распределения данных.

Проведем внутригрупповой корреляционный анализ:

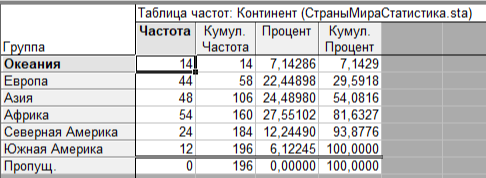


В группе Есть нефть?: Корреляция между населением и плотностью населения (-0,001) является низкой и статистически не значимой. Корреляция между Плотностью населения и Площадью (-0,097) является статистически значимой, но всё такой же низкой.

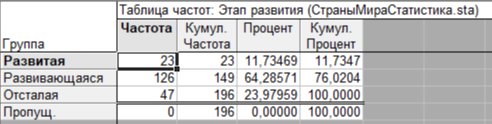


**4 Частотный анализ**

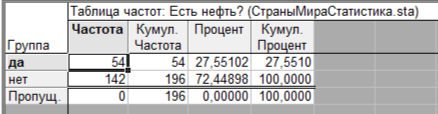
**4.1 Таблицы частот**

****

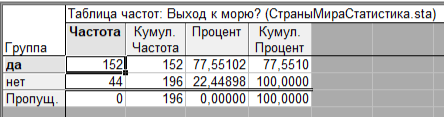
Данная таблица частот построена для категориальной переменной Континент. В таблице для всех континентов указаны слева направо: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты. Можно сделать вывод, что больше всего стран находится в Африке, Азии и Европе.

****

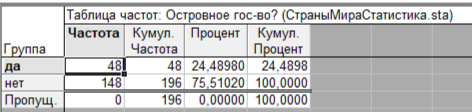
Данная таблица частот построена для категориальной переменной Этап развития. В таблице указаны слева направо: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты. Можно сделать вывод, что больше всего развивающихся стран.



Данная таблица частот построена для категориальной переменной Есть нефть?. В таблице указаны слева направо: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты. Можно сделать вывод, что больше стран, у которых нет нефти.

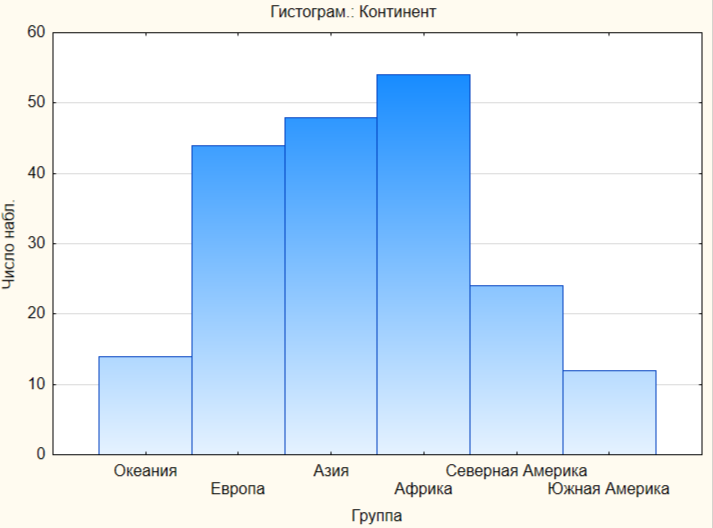


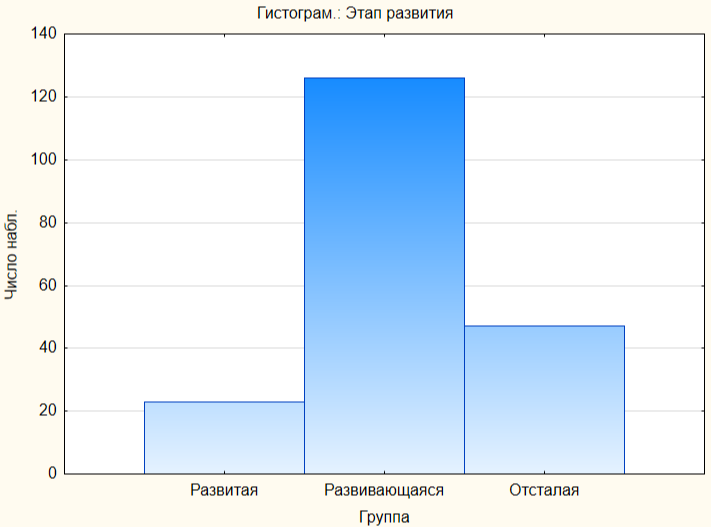
Данная таблица частот построена для категориальной переменной Выход к морю?. В таблице указаны слева направо: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты. Можно сделать вывод, что больше стран, у которых есть выход к мировому океану.

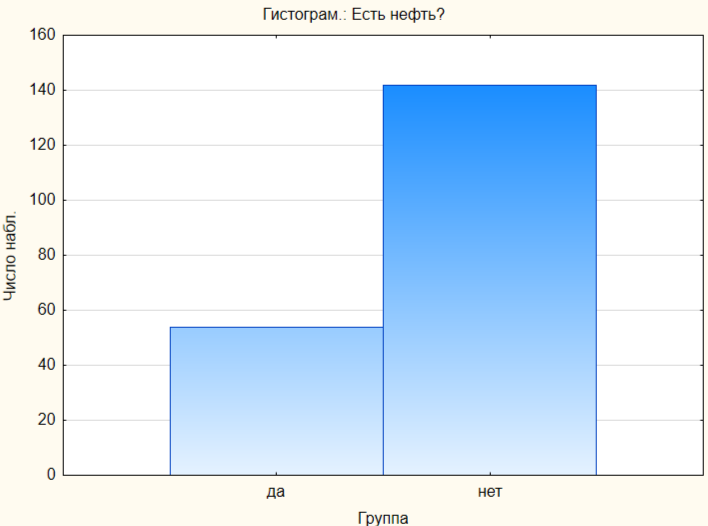


Данная таблица частот построена для категориальной переменной Островное гос-во?. В таблице указаны слева направо: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты. Можно сделать вывод, что больше стран, которые располагаются на материке, нежели островных государств.

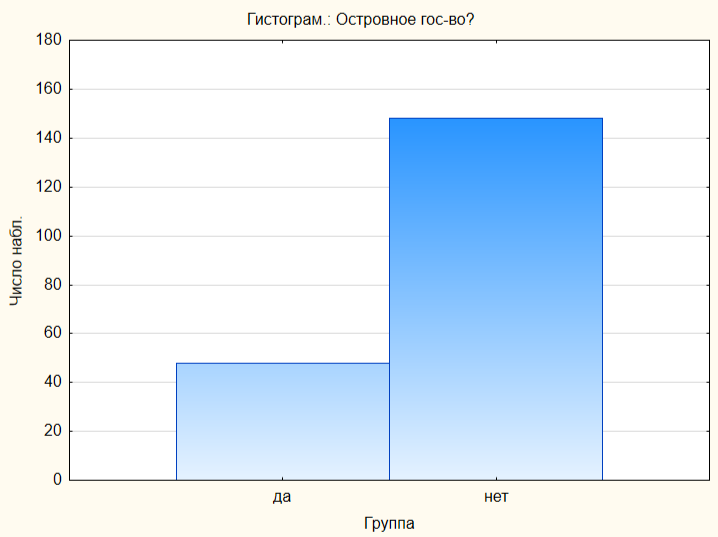
Гистограммы соответствующих категориальных переменных:





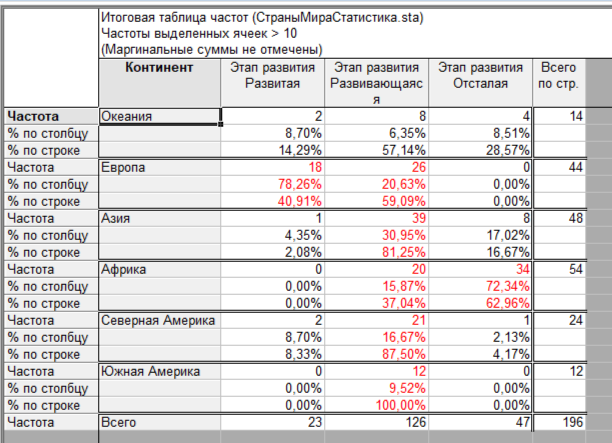






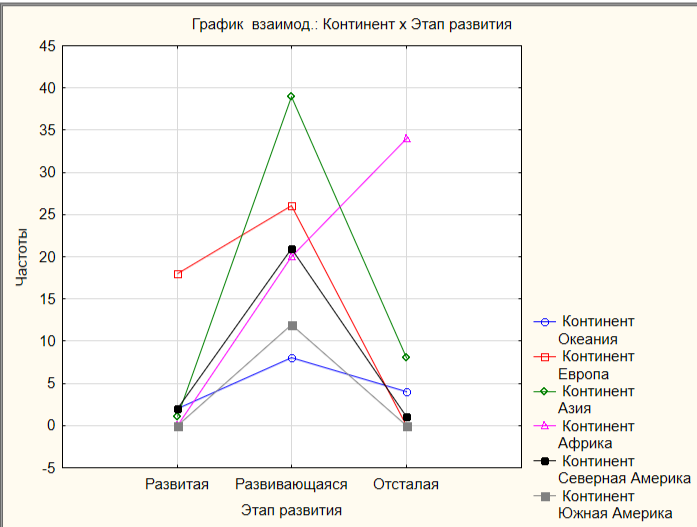
**4.2 Таблицы кросстабуляции**

Таблицы кросстабуляции объединяют несколько таблиц частот.

****

В данной таблице показаны частоты совокупности двух категориальных переменных *Континенты* и *Этап развития*.

Графики взаимодействия частот:

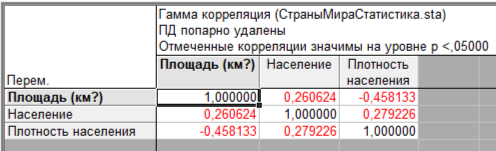
****

По таблице и графику можно сделать выводы, что в Европе больше всего развитых стран, в Азии больше развивающихся стран, а Африка лидер по отсталым странам.

**5 Корреляционный анализ**

Нулевая гипотеза – «между переменными корреляция статистически не значима». Коэффициент Спирмена используется, если закон распределения переменных неизвестен или не является нормальным.

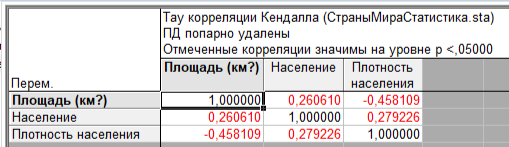
Проведем ранговые корреляции для переменных площади, населения и его плотности:



Так как для переменных переменных площади, населения и его плотности p<0.05, то нулевая гипотеза отвергается, то есть корреляция между переменными статистически значима.

Коэффициент Кендалла используется, если хотя бы одна переменная качественная.

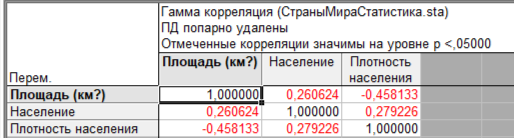
Проведем Тау корреляции для переменных площади, населения и его плотности:



Так как p<0.05, то нулевая гипотеза отклоняется, то есть корреляция между переменными статистически значима.

Коэффициент Гамма используют, если переменные содержат много повторяющихся значений.

Проведем Гамма корреляции для переменных площади, населения и его плотности:



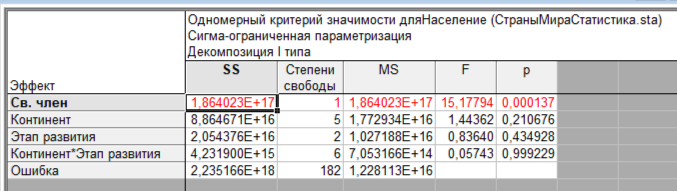
Так как p<0.05, то нулевая гипотеза отвергается, то есть корреляция между переменными статистически значима.

**6 Дисперсионный анализ**

Дисперсионный анализ применяется**,** если сравниваем подгруппы по двум  группирующим переменным.

Дисперсионный анализ является наиболее общим методом сравнения средних. В дисперсионном анализе можно исследовать зависимость количественного признака (зависимой переменной) от одного или нескольких качественных признаков (факторов).

В качестве зависимой переменной выбран население в качестве категориальных континент, этап развития.



Получили таблицу в строках которой будут указаны названия эффектов, значения критерия Уилкса, критерия Фишера дисперсионного анализа и в последнем столбце отображены уровни значимости критерия Фишера. Т.к. для первого р<0,05, то это означает, что при разбиении на группы по данному эффекту наверняка будут группы со статистически значимыми отличиями средних.

**7 Линейное многомерное моделирование взаимосвязей**

Построим зависимость между площадью и населением с плотностью населения.

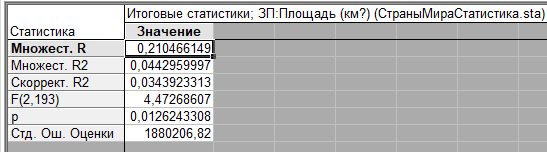




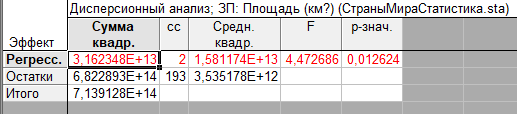
Таблица содержит стандартизованные Бета и не стандартизованные (*B*) регрессионные коэффициенты (веса), их стандартные ошибки и уровни значимости. Коэффициенты *Beta* оцениваются по стандартизованным данным, имеющим выборочное среднее, равное 0 и стандартное отклонение, равное 1. Поэтому величины *Beta* позволяет сравнить вклады каждого предиктора в предсказание отклика.Отрицательный знак коэффициентов при этих переменных означает, что с увеличением одной переменной – парная ей с отрицательной корреляцией падает. Положительный знак коэффициента при переменнойозначает, что с ее увеличением, у увеличивается и парная. Коэффициенты уравнения регрессии *b1*, *b2* и свободный член статистически значимы при уровне значимости *p >* 0,05

Частная корреляция:

****

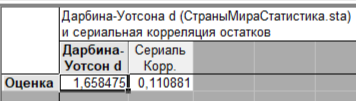
Частные коэффициенты корреляции показывают степень влияния одного предиктора на отклик в предположении, что остальные предикторы закреплены на постоянном уровне, т.е. контролируется их влияние на отклик. Частные коэффициенты корреляции так же, как и стандартизованные коэффициенты*Beta* позволяют провести ранжирование предикторов по степени их влияния на отклик.

Таблица с результатами дисперсионного анализа для проверки гипотезы о значимости уравнения регрессии при помощи критерия Фишера:

**

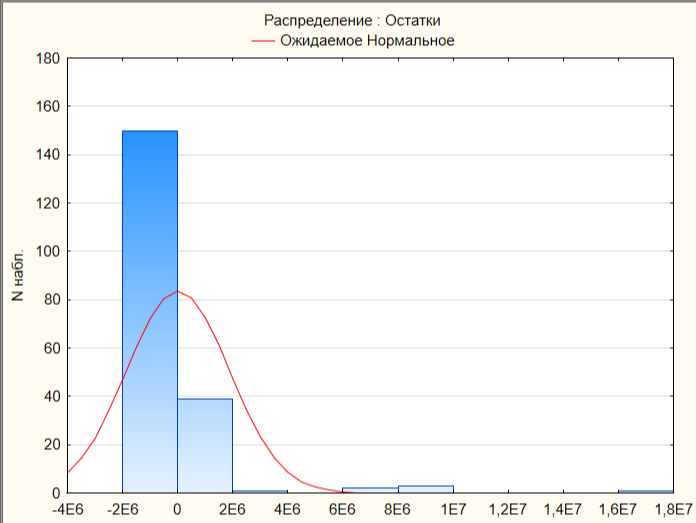
Первая строка из этой таблицы присутствует в информационной части окна с результатами анализа.

статистика Дарбина **–** Уотсона:



Эта статистика характеризует наличие или отсутствие сериальной корреляции (зависимости) между остатками для соседних наблюдений. Существование сериальной корреляции может служить доказательством зависимости наблюдений в файле данных. Из таблицы видно, что статистика Дарбина **–** Уотсона имеет небольшое значение (≈1,66) при средней сериальной корреляции (≈ 0,111). Это свидетельствует о некоторой зависимости наблюдений, следовательно, можно говорить о недостаточной устойчивости некоторых значений коэффициентов регрессии, а значит о невысокой адекватности модели изучаемому процессу.

Гистограмма остатков:



По графику можно сделать вывод, что модель практически адекватная, т.к. распределение остатков практически соответствует нормальному закону со средним значением 0.

**8 Нелинейное многомерное моделирование взаимосвязей**

**8.1 Fixed Nonlinear Regression**

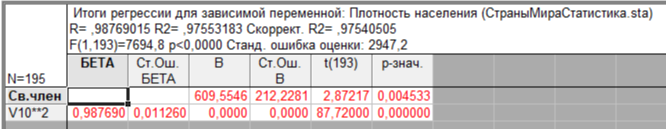
Этот модуль реализует множественный линейный регрессионный анализ с линеаризованной моделью.

Проведем множественную нелинейную регрессию площади и населения от плотности населения.



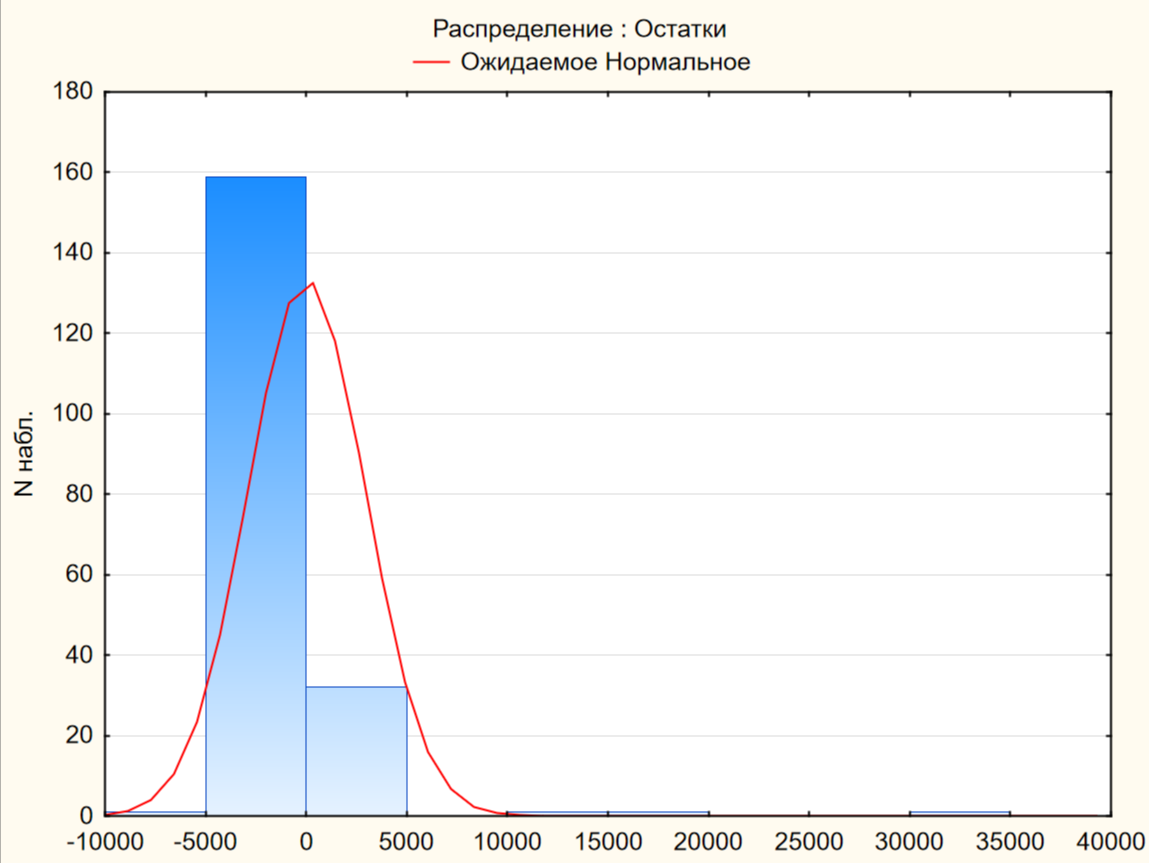
Как видно, предиктор модели V10\*\*3 статистически значим, при статистически значимом уравнении регрессии. Остальные предикторы модели статистически незначимы, при статистически значимом уравнении регрессии.

Проведем множественную нелинейную регрессию с методом пошагового исключением.



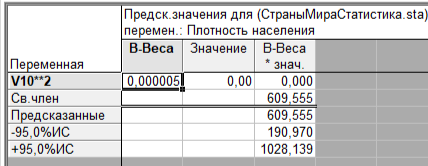
Площадь и население = 609,5545

Для проверки адекватности модели построим гистограмму остатков.



Видно, что распределение остатков практически соответствуют нормальному, поэтому модель можно назвать адекватной.

Предсказание значения для автомобиля с engine\_power = 4,3, mileage = 15000:

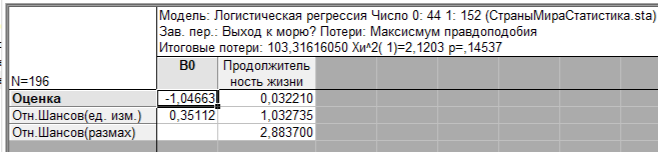


Из таблицы видно, что предсказанная стоимость равна 610.

**8.2 Логит регрессия**

Таблица оценки параметров.

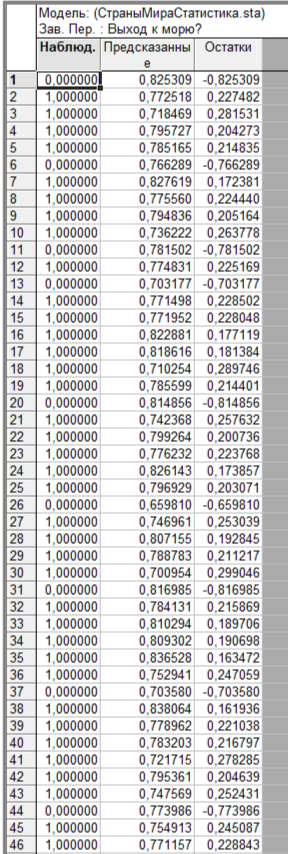
Построим Логит модель зависимости бинарной переменной Выход к морю? от продолжительности жизни. Метод оценивания – Хука-Дживиса.



Регрессия не значима, так как p>0.05.

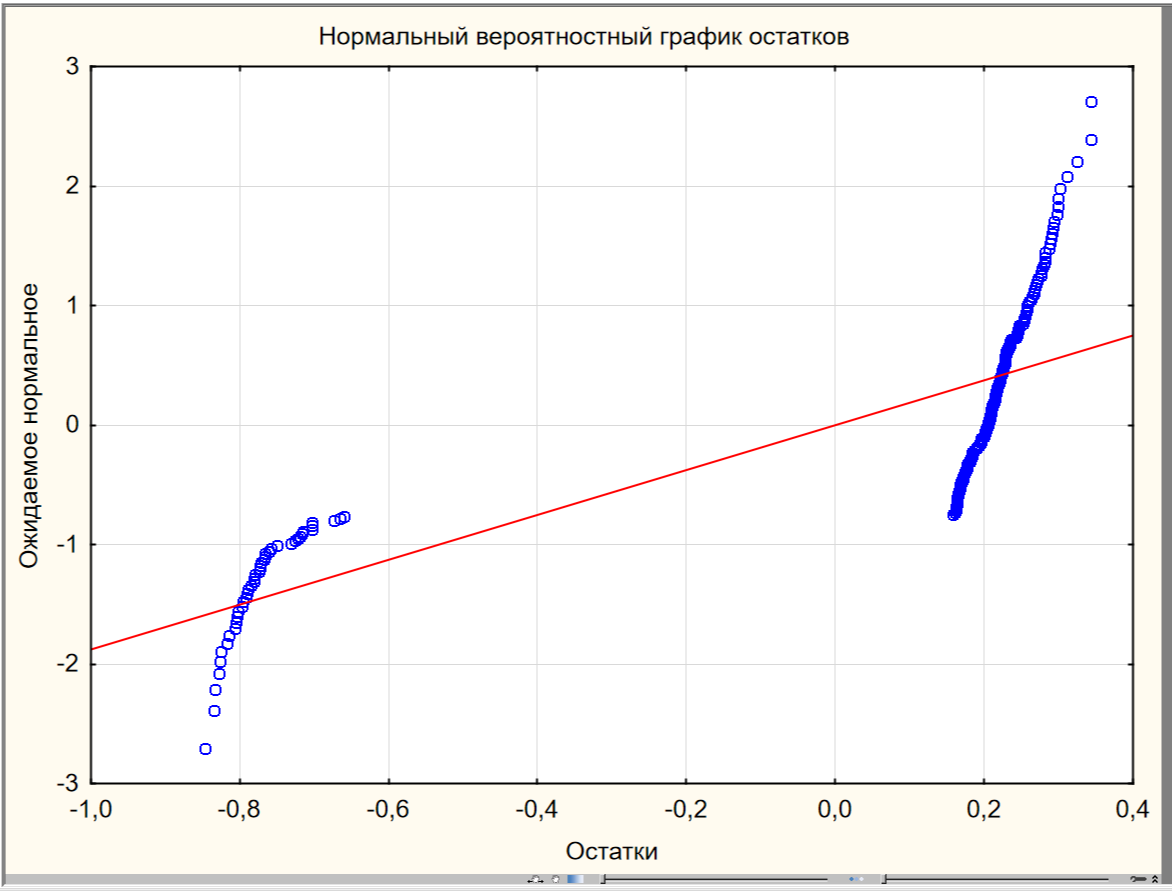
Выход к морю? = exp(-1,05) / (1+exp(0,35112+1,03\*продолжительность жизни))

Таблица с Наблюдениями, предсказаниями и остатками:



Модель лучше предсказывает страны относящиеся к Выход к морю?

Нормальный вероятностный график остатков:



Данный график показывает, что распределение остатков (ошибок) близко к нормальному.

**8.3 Экспоненциальная регрессия**

Построим экспоненциальную модель зависимости плотности населения от населения:

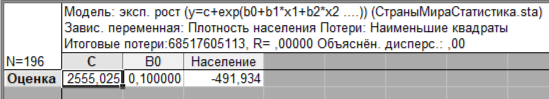
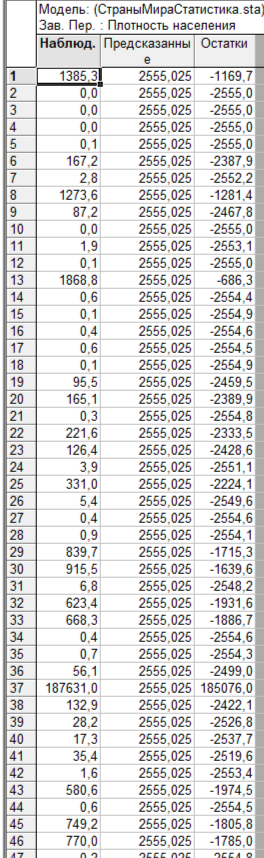
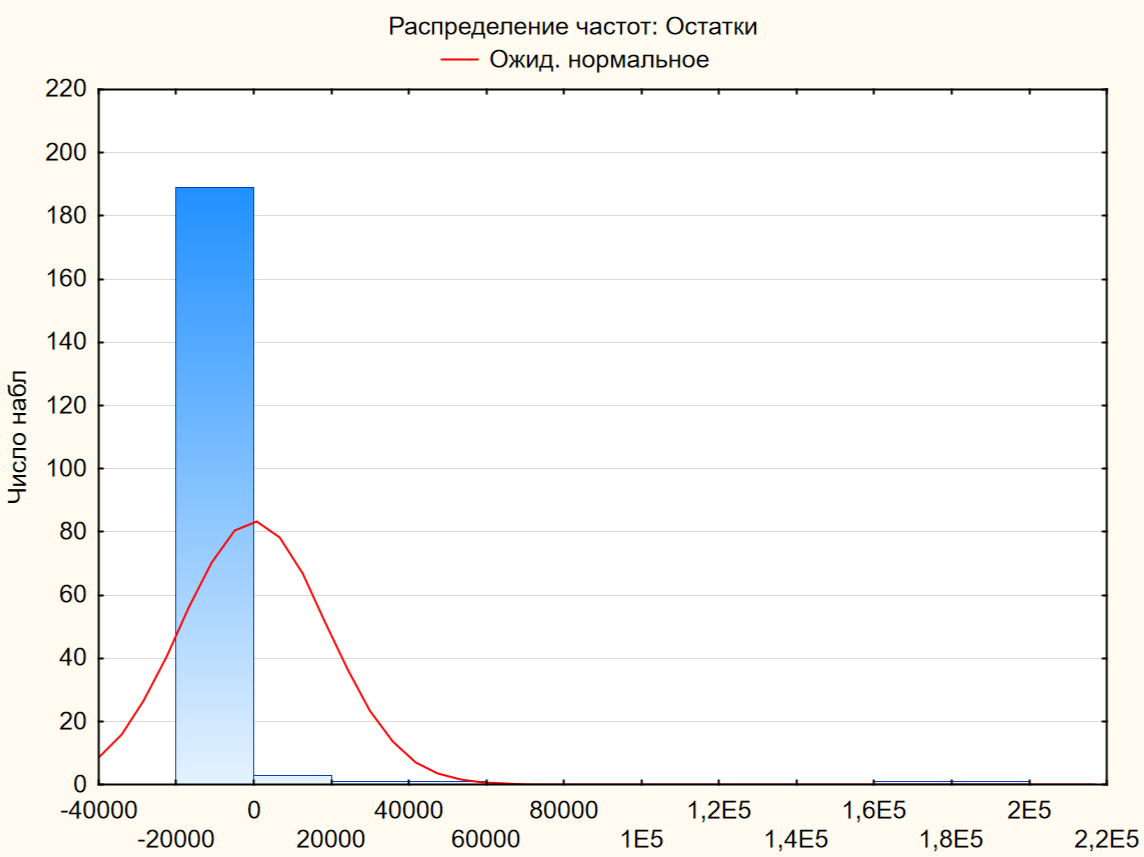


Таблица предсказанных и наблюдаемых значений:

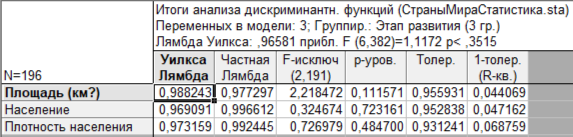




По графикам остатков видно, что модель является адекватной.

**9 Дискриминантный анализ**

Дискриминантный анализ проводим по группирующей переменной, полученной в результате кластеризации не менее, чем на 3 группы.

****

Чем больше значение λ, тем более желательно присутствие

этой переменной в процедуре дискриминации. Все переменные имеют высокое значение Уилкса лямбда примерно равное 0.92, что говорит о их одинаково высоком влиянии в процедуре дискриминации.

В первом столбце таблицы приведены значения *Wilks Lambda*, являющиеся результатом исключения соответствующей переменной из модели. Чем больше значение λ, тем более желательно присутствие этой переменной в  процедуре дискриминации.

Значение *Partial Lambda* (частная лямбда) есть отношение общей лямбда Уилкса к лямбде Уилкса после исключения этой переменной. Частная лямбда характеризует единичный вклад соответствующей переменной в разделительную силу модели. Чем меньше частная лямбда Уилкса, тем больше вклад в общую дискриминацию. У всех переменных одинаковое значение частной Лямбда, что говорит о их одинаковом влиянии на общую дискриминацию.

*F-removeтпп* (*F-исключить*) – это значения *F-критерия*, связанные с соответствующей частной лямбда Уилкса. Значения *p-level* – это уровни значимости критериев *F-remove*.

      Толерантность (*Toler*) определяется как *1-R2,* где *R2* – это коэффициент множественной корреляции данной переменной со всеми другими переменными в модели. Как уже отмечалось, толерантность является мерой избыточности переменной в модели. Чем меньше значение толерантности, тем избыточнее переменная в модели.

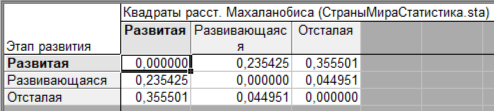
*F-remove* (*F-исключить*) – это значения *F-критерия*, связанные с соответствующей частной лямбда Уилкса. Значения *p-level* – это уровни значимости критериев *F-remove*.

      Толерантность (*Toler*) определяется как *1-R2,* где *R2* – это коэффициент множественной корреляции данной переменной со всеми другими переменными в модели. Как уже отмечалось, толерантность является мерой избыточности переменной в модели. Чем меньше значение толерантности, тем избыточнее переменная в модели.

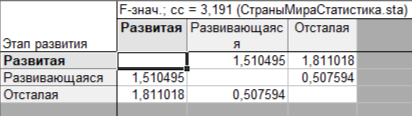
*F-remove* (*F-исключить*) – это значения *F-критерия*, связанные с соответствующей частной лямбда Уилкса. Значения *p-level* – это уровни значимости критериев *F-remove*.

Толерантность (*Toler*) определяется как *1-R2,* где *R2* – это коэффициент множественной корреляции данной переменной со всеми другими переменными в модели. Как уже отмечалось, толерантность является мерой избыточности переменной в модели. Чем меньше значение толерантности, тем избыточнее переменная в модели.

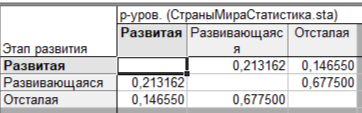
Расстояние между группами:



Квадраты расстояния Махаланобиса.

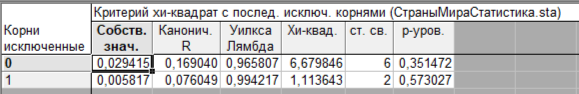


F значения говорят о том, что данные статистически значимы.

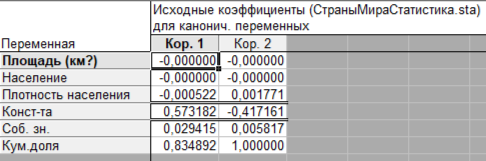
****

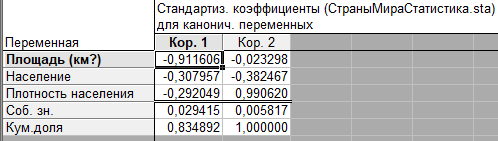
p – уровень <0,05 значит данные статистически значимы.

Канонический анализ, критерий Хи-квадрат:

****

Первая строка дает критерий значимости для всех корней. Вторая строка содержит значимость корней, оставшихся после удаления первого корня и т.д. Как видно из таблицы, вторая дискриминантные функции статистически незначима.

****

****

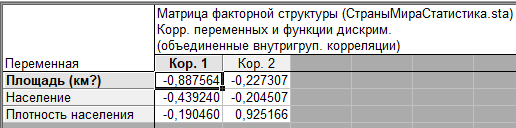
В первой таблице даны исходные (не стандартизованные) коэффициенты дискриминантных функций. Эти коэффициенты могут быть использованы для вычисления значений канонических переменных для каждого наблюдения. Во второй таблице приведены стандартизованные коэффициенты дискриминантных функций.

Эти коэффициенты, основанные на стандартизованных переменных, принадлежат к одной и то же шкале измерений (абсолютной), поэтому их можно сравнивать, чтобы определить величины и направления вкладов переменных в каждую каноническую функцию.

Из таблицы видно, что наибольший вклад в дискриминантную функцию 1 вносит переменная *Площадь (-0,912)*, во 2ю функцию переменная Плотность населения (0,991).

Так же в таблицах приведены собственные значения для дискриминантной функции и кумулятивная доля объясненной дисперсии, накопленной функцией.

Факторная структура:



В таблице приведены объединенные внутригрупповые корреляции переменных с соответствующей дискриминантной функцией. Эти корреляции называют еще структурными коэффициентами. Обычно структурные коэффициенты используют для содержательной интерпретации функций, в отличие от коэффициентов дискриминантной функции, которые обозначают вклад каждой переменной в функции. У переменных *Плотность населения (0,991), Население (-0,382), Площадь (0,032)* наибольшие корреляции с 2й дискриминантной функцией.

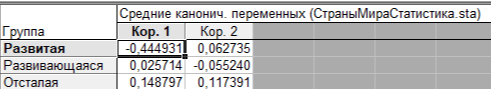
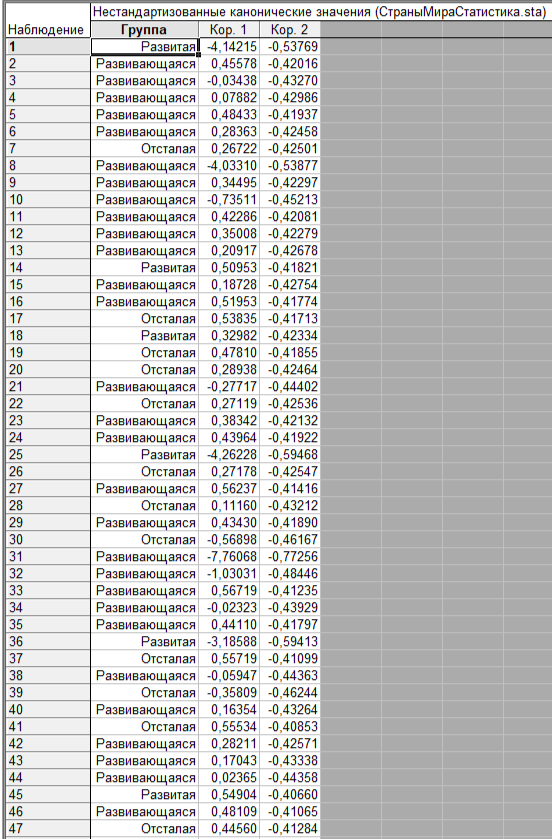
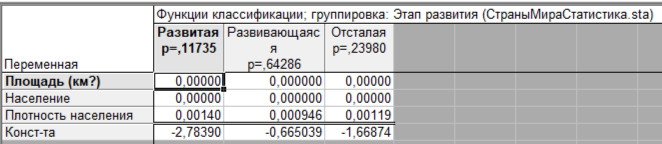


Таблица со средними значениями для дискриминантных функций, которые позволяют определить группы, лучше всего идентифицируемые конкретной дискриминантной функцией. Группа Развитая (-0,445), Отсталая (0,149) лучше всего идентифицируемы с 1й дискриминантной функцией, а Развивающаяся (-0,055) – с 2й дискриминантной функцией.



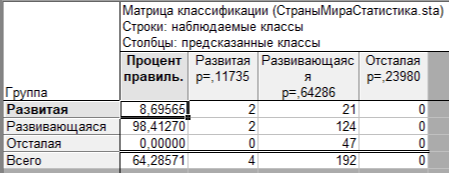
Данная таблица со значениями дискриминантных функций для каждого наблюдения. По таблице трудно судить о результатах разделения программой наблюдений по группам.



Функции классификации – это линейные функции, которые вычисляются для каждой группы и могут быть использованы для классификации наблюдений.

Наблюдение приписывают той группе, для которой классификационная функция имеет наибольшее значение. В таблице приведены коэффициенты и свободные члены при переменных линейных функций. Например, классификационное уравнение для группы low имеет вид:

Automatic = 0\*Площадь + 0\*Население + 0,0014\*Плотность населения – 2,7839



Матрица содержит информацию о количестве и проценте корректно классифицированных наблюдений в каждой группе. Строки матрицы – действительные наблюдения, столбцы – предсказанные наблюдения.

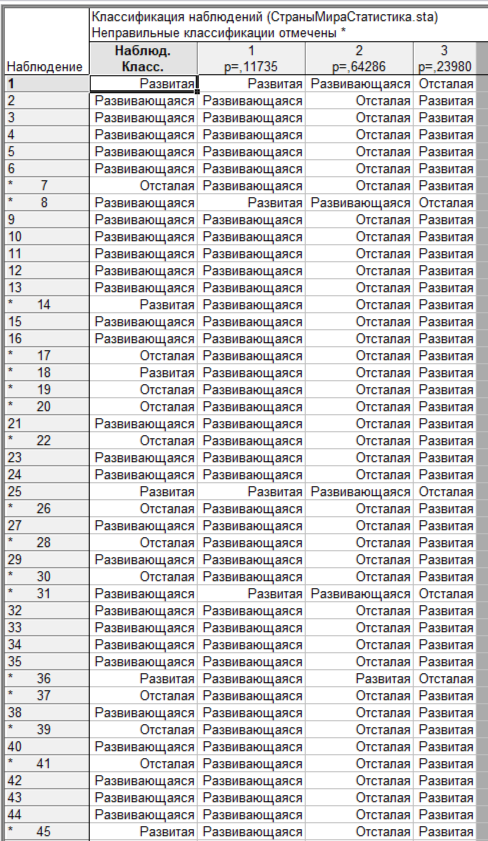
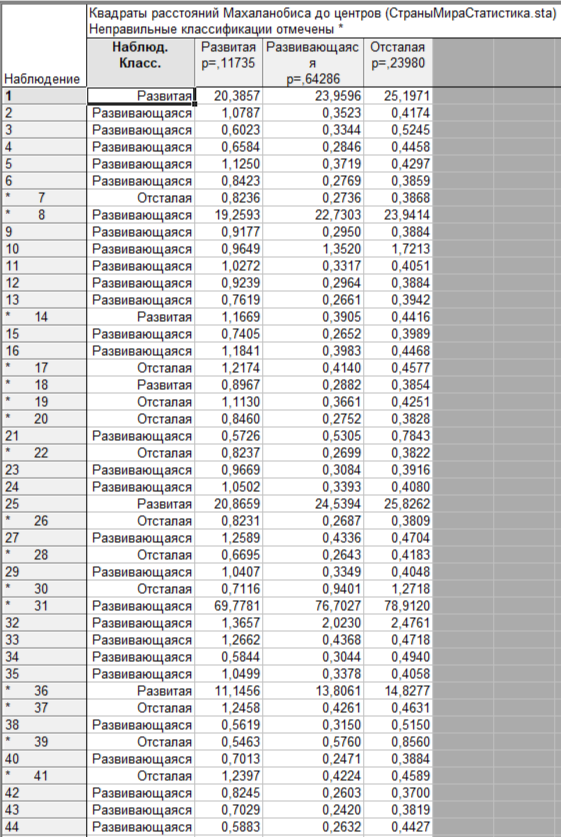


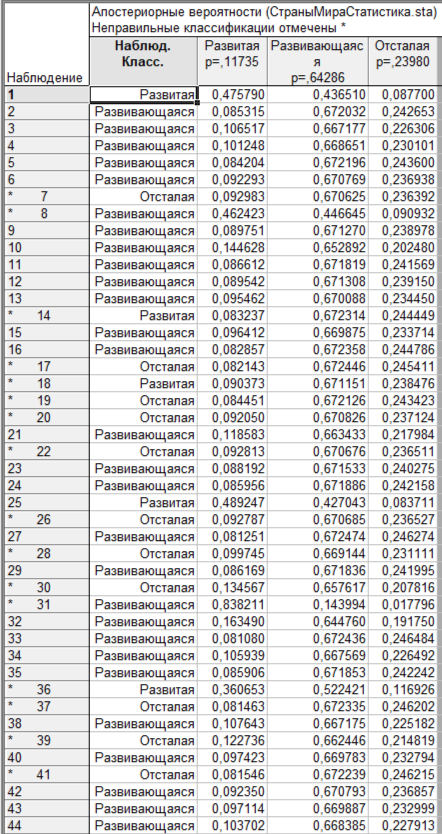
Таблица классификации для каждого наблюдения. Классификации упорядочены по первому и второму выбору. Столбец 1 содержит первый классификационный выбор, т. е. группу, для которой соответствующее наблюдение имеет наивысшую апостериорную вероятность и наибольшее значение классификационной функции. Наблюдения, которые не удалось правильно классифицировать, помечены \*.

Таблица квадратов расстояний Махалонобиса каждого наблюдения от центроида группы



Эти расстояния аналогичны квадратам евклидовых расстояний, но учитывают корреляции между переменными в модели. Наблюдение приписывают группе, к которой оно ближе всего. Наблюдения, которые не удалось правильно классифицировать, также помечены \*.

Таблица апостериорных вероятностей:



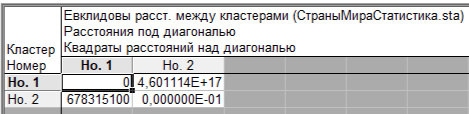
В таблице каждому наблюдению будет поставлена в соответствие вероятность принадлежности к группам. Эта вероятность определяется посредством расстояний Махалонобиса и априорных вероятностей. Чем дальше наблюдение расположено от центра группы, тем менее вероятно, что оно принадлежит этой группе. Наблюдение приписывают той группе, для которой имеется наибольшая апостериорная вероятность классификации.

**10** **Кластерный анализ**

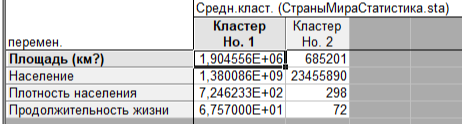
**10.1 Метод к-средних**

Кластерный анализ позволяет разбить множество исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры.

Попробуем разбить автомобили на кластеры, но до этого надо стандартизировать переменные. Варьируя количеством кластеров и исключая переменные, получили следующие результаты



В данной таблице под диагональю указаны расстояния между кластерами, над диагональю квадраты расстояний.

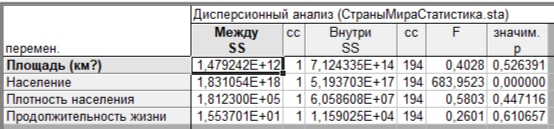


В данной таблице указаны средние значения переменных для каждого из кластеров. У переменной Площадь значение 1,904556E+06 для первого кластера и 685201 для второго, у переменной Население 1,380086E+09

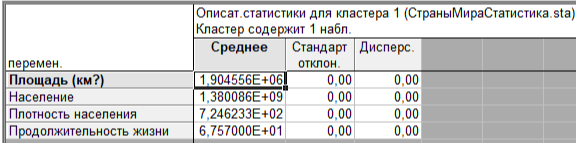
для первого кластера и 23455890 для второго. У переменной Плотность населения 7,246233E+02 для первого и 298для второго, у переменной Продолжительность жизни 6,757000E+01 для первого кластера и 72

для второго.

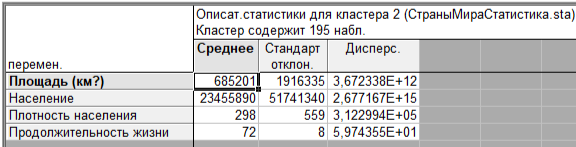
Таблица дисперсионного анализа



В таблице приведены значения межгрупповых *(Между SS)* и внутригрупповых *(Внутри SS)* дисперсий признаков. Чем меньше значение внутригрупповой дисперсии и больше значение межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру и тем «качественнее» кластеризация. Параметры *F* и *p* также характеризуют вклад признака в разделение объектов на группы. Лучшей кластеризации соответствуют большие значения первого и меньшие значения второго параметра. Признаки с большими значениями *р* (например, больше 0,05) можно из процедуры кластеризации исключить.

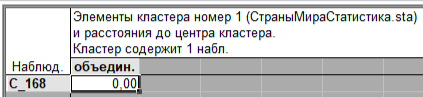


В данной таблице указаны описательные статистики по переменным для первого кластера.

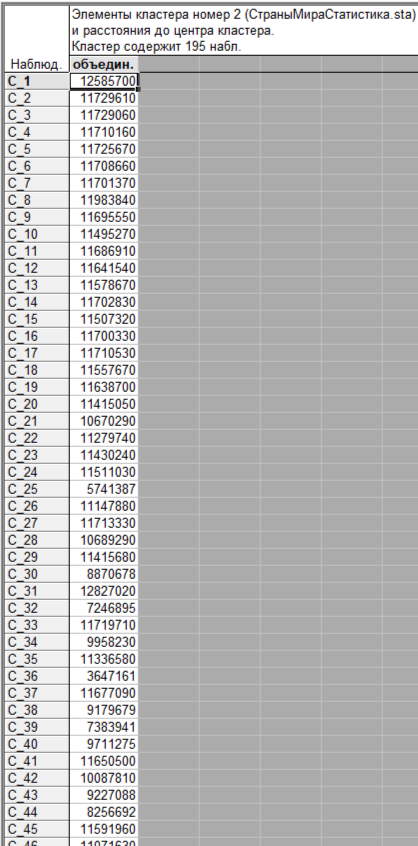


В данной таблице указаны описательные статистики по переменным для второго кластера.

Построим таблицу принадлежности объектов к кластерам



В таблице указаны элементы кластера 1 с расстоянием до центра кластера, например, элемент С\_168 принадлежит кластеру 1 с расстоянием до центра 0, что логично, так как он один.



В таблице указаны элементы кластера 2 с расстоянием до центра кластера, например, элемент С\_1 принадлежит кластеру 2 с расстоянием до центра 12585700.

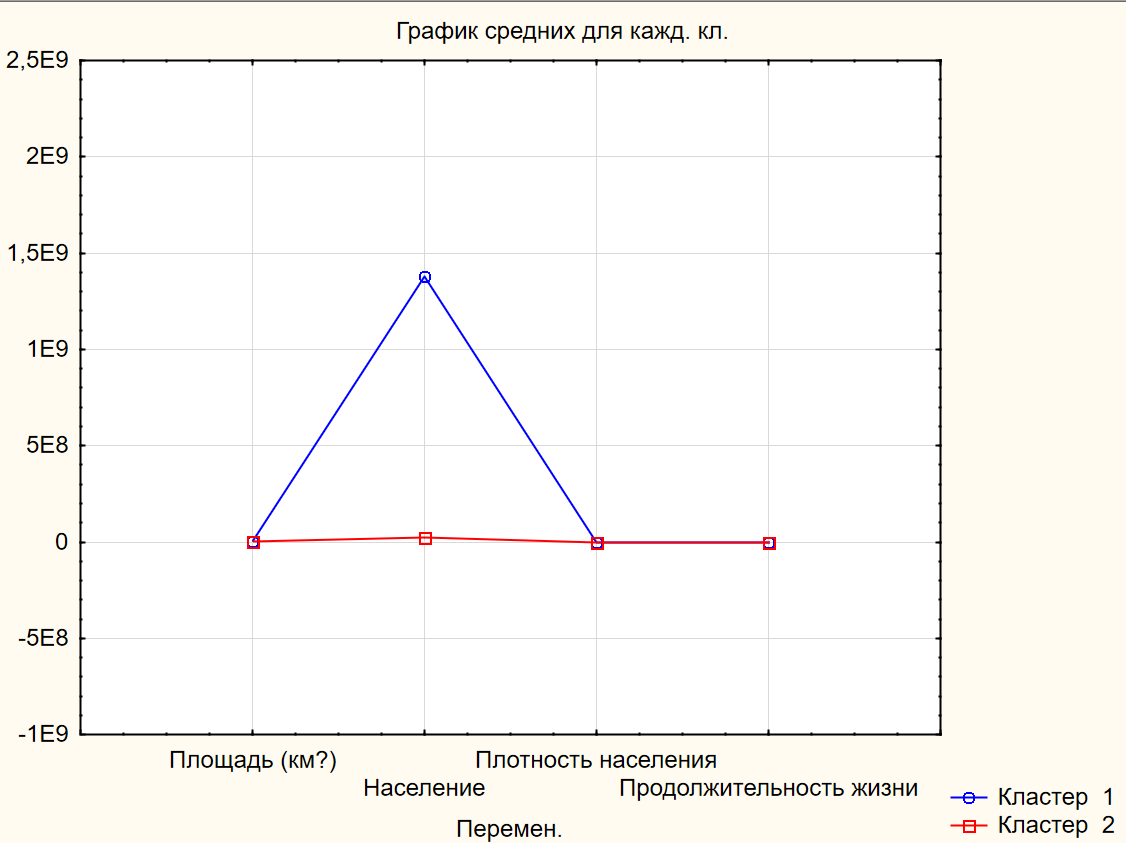


График позволяет просмотреть средние значения для каждого кластера на линейном графике. Видно, что в кластерах 1 и 2 средние параметры *население* значительно отличаются.Это свидетельствует о удовлетворительном разбиении на группы.

**10.2 Древовидная кластеризация**

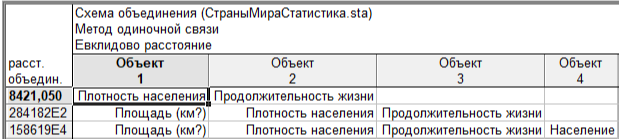
Проведем древовиднуюкластеризацию по объектам.

Горизонтальная древовидная диаграмма имеет вид:



На диаграмме видно, что по мере понижения порога (изменение расстояния объединения), относящегося к решению об объединении объектов в один кластер, все большее и большее число объектов связывается, и объединяются все больше кластеров.

Правило объединения в кластеры показано в следующей таблице:



По данной таблице видно компоновка объектов в кластеры с расстояниями объединений.

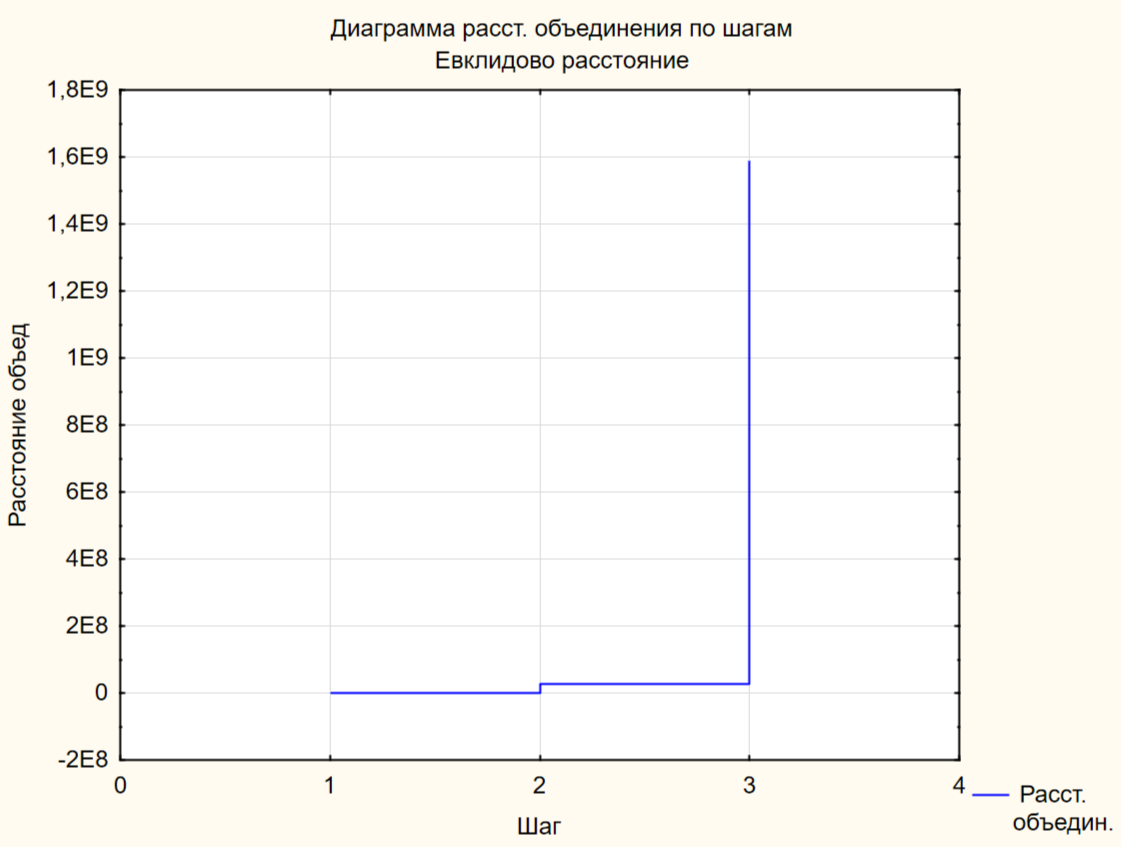
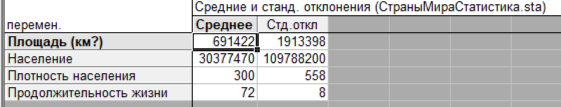
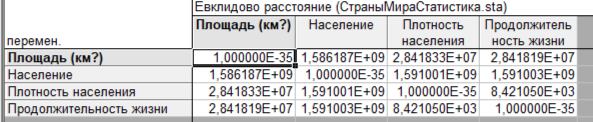


Таблица средних и стандартных отклонений:



В таблице можно увидеть средние и стандартные отклонения.

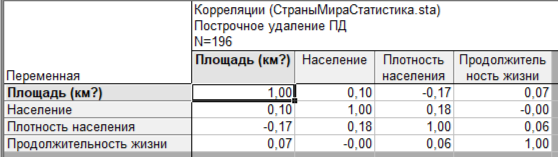
Таблица Евклидовых расстояний



В таблице указаны евклидовы расстояния между элементами.

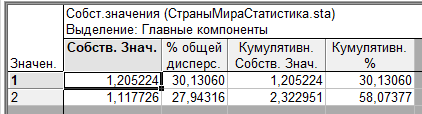
**1****1 Факторный анализ**

Главная цель факторного анализа это сокращение числа переменных и определение структуры взаимосвязей между переменными, т.е. классификация переменных.



Из таблицы видны значения корреляций между переменными. Анализ проведем по переменным данным переменным:

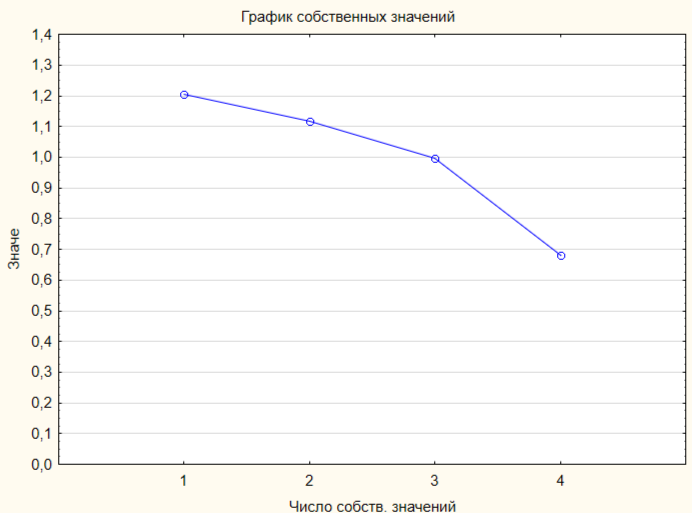
Просмотрим собственные значения факторов:



Во втором столбце приведены дисперсии выделенных факторов – собственные числа, в третьем – процент от общей дисперсии. Как видно, первый фактор объясняет 30% общей дисперсии, второй – 28%. Необходимо понять, сколько факторов следует оставить.

По критерию Кайзера можем отобрать только факторы с собственными значениями, большими 1. Из таблицы видно, что на основе данного критерия выделяются оба фактора.

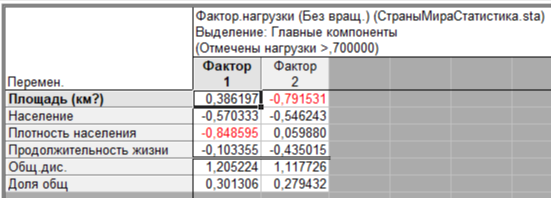
Критерий каменистой осыпи является графическим:



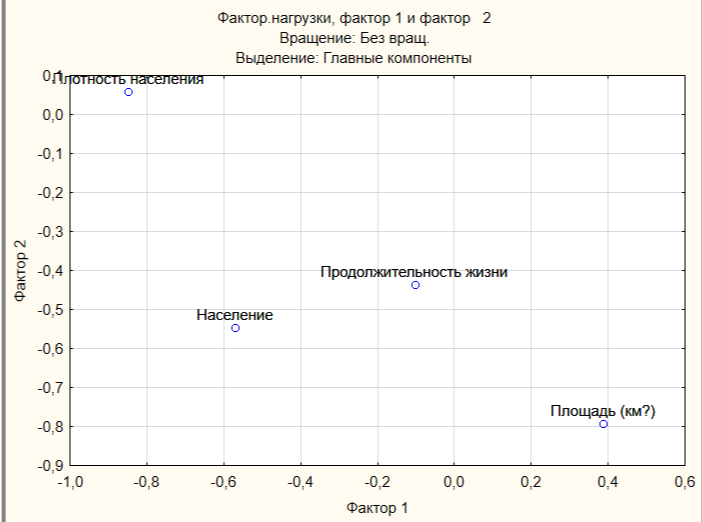
Надо выбрать такое место на графике, где убывание собственных значений слева направо максимально замедляется. Из графика видно, что в соответствии с этим критерием можно попытаться выделить 2 фактора.

Предположим, что число факторов неизвестно.

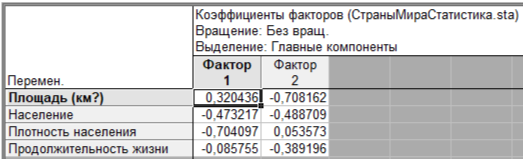
Корреляции между переменными и выделенными факторами:



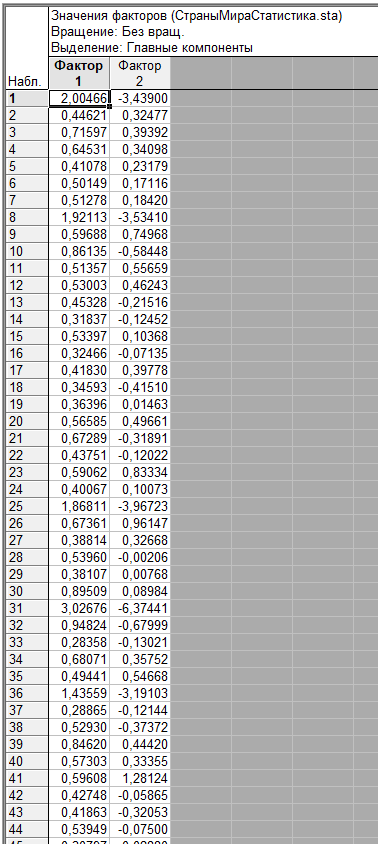
Из таблицы видно, что первому и второму факторам соответствуют высокие коэффициенты корреляции.



Этот график иллюстрирует соотношение между факторами и группами переменных. Коэффициенты уравнений регрессий, по которым программа посчитает значения факторов для каждого наблюдения, представлены в таблице:



Значения факторов для каждого автомобиля представлены в таблице:



По данным из этой таблицы можно судить об отношении страны к соответствующему фактору. Положительное значение фактора соответствует позитивному отношению, а отрицательное – негативному.