Оглавление

[1. Постановка задачи 3](#_Toc452976423)

[1.1. Формулировка прикладной задачи 3](#_Toc452976424)

[1.2. Потенциальные потребители решения 3](#_Toc452976425)

[1.3. Основные гипотезы 4](#_Toc452976426)

[1.4. Основные источники данных 4](#_Toc452976427)

[1.5. Публикации 4](#_Toc452976428)

[2. Предварительный анализ собранных данных 5](#_Toc452976429)

[2.1. Анализ особенностей данных 5](#_Toc452976430)

[2.1.1. Графический анализ данных и описательная статистика 6](#_Toc452976431)

[2.1.1.1. Количество преступлений 6](#_Toc452976432)

[2.1.1.2. Численность населения 7](#_Toc452976433)

[2.1.1.3. Процент сбережений от доходов 8](#_Toc452976434)

[2.1.1.4. Количество безработных 8](#_Toc452976435)

[2.1.1.5. Количество посещений музеев на 1000 человек 9](#_Toc452976436)

[2.1.1.6. Численность населения с доходами ниже прожиточного минимума 10](#_Toc452976437)

[2.1.1.7. Количество выпускников на ВУЗ на 1000 человек 10](#_Toc452976438)

[2.1.1.8. Плотность населения 11](#_Toc452976439)

[2.1.1.9. Качество жизни населения 15](#_Toc452976440)

[2.1.2. Пропущенные значения и выбросы 20](#_Toc452976441)

[2.2. Анализ статистической связи 20](#_Toc452976442)

[2.2.1. Графический анализ 20](#_Toc452976443)

[2.2.2. Корреляционная матрица 24](#_Toc452976444)

[2.3. Вывод из предварительного анализа данных 25](#_Toc452976445)

[3. Построение модели 26](#_Toc452976446)

[3.1. Базовая модель 26](#_Toc452976447)

[3.1.1. Связь структуры базовой модели с основными гипотезами исследования 27](#_Toc452976448)

[3.1.2. Связь структуры базовой модели с результатами предварительного анализа данных 27](#_Toc452976449)

[3.1.3. Оценивание базовой модели 28](#_Toc452976450)

[3.1.2 Проверка гипотез для базовой модели 30](#_Toc452976451)

[3.2. Оптимизированная линейная регрессия 31](#_Toc452976452)

[3.2.1. Тесты на гетероскедастичность и мультиколлинеарность 32](#_Toc452976453)

[3.2.2 Проверка гипотез для оптимизированной модели 33](#_Toc452976454)

[3.3. Сравнение начальной модели с оптимизированной 33](#_Toc452976455)

[3.3.1. Модифицированный коэффициент детерминации 33](#_Toc452976456)

[3.3.2. Критерий Акаике (AIC) 34](#_Toc452976457)

[3.3.3. Критерий Шварца (BIC) 34](#_Toc452976458)

[3.4 Вывод 34](#_Toc452976459)

[4. Построение доверительных интервалов и проведение тестов на выбросы 34](#_Toc452976460)

[4.1. Доверительные интервалы 34](#_Toc452976461)

[4.2. Тесты на выбросы 35](#_Toc452976462)

[5. Заключение 36](#_Toc452976463)

# Постановка задачи

## Формулировка прикладной задачи

Проблему, которую может помочь решить наше эконометрическое исследования, мы обозначили следующим образом: «Определить, какие социальные сферы следует финансировать государству и в какой пропорции, чтобы эффективнее всего способствовать снижению уровня преступности». Для этого мы проанализируем следующие переменные:

* Численность населения – измеряется количеством человек, проживающих в конкретном регионе в 2014 году; количественная независимая переменная;
* Процент сберегаемых средств – измеряется процентом от дохода населения в конкретном регионе в 2014 году, который не тратится ни на какие нужды; количественная независимая переменная;
* Количество безработных людей – измеряется количеством человек, неустроенных на работу в конкретном регионе в 2014 году; количественная независимая переменная;
* Численность населения с доходами ниже прожиточного минимума – измеряется количеством человек, получающих доход ниже прожиточного минимума в конкретном регионе в 2014 году; количественная независимая переменная;
* Плотность населения – измеряется одним из пяти текстовых значений, подробнее о том, по какой логике присваивались значения будет сказано позднее; номинальная независимая переменная, допустимые значения – «маленькая», «средняя», «выше среднего», «большая»;
* Количество выпускников на ВУЗ на 1000 человек – измеряется количеством человек, выпущенных всеми ВУЗами конкретного региона в 2014 году, поделенным на количество ВУЗов в конкретном регионе и приведенным к количеству на 1000 человек; количественная независимая переменная;
* Количество посещений музеев на 1000 человек – измеряется количеством посещений музеев в конкретном регионе в 2014 году, приведенным к количеству на 1000 человек; количественная независимая переменная;
* Количество совершенных преступлений – измеряется количеством преступлений всех возможных видов, совершенных в конкретном регионе в 2014 году; количественная зависимая переменная;
* Рейтинг качества жизни – измеряется одним из двух текстовых значений, подробнее о том, по какой логике присваивались значения будет сказано позднее; номинальная зависимая переменная, допустимые значения – «выше среднего», «ниже среднего»

## Потенциальные потребители решения

Очевидно, что главным потребителем результатов нашего исследования может быть правительство РФ. С помощью построенной модели и найденных зависимостей между совершенными преступлениями и другими социальными факторами можно будет определить «драйверы» преступности и принимать меры, направленные на снижения этих «драйверов». Также возможным потребителем исследований могут стать правоохранительные органы так как особое внимание к потенциально рискованным группам граждан может предотвратить серьезные преступления. Однако для более точных результатов необходимо использовать данные максимально приближенные к реальным и использовать программные средства, которые способны спрогнозировать развитие ситуации при имеющихся параметрах.

## Основные гипотезы

1. Первая, самая очевидная гипотеза заключается в наличии прямой зависимости между количеством совершенных преступлений и количеством населения в конкретном субъекте федерации.
2. Существует зависимость между процентом сбережений от получаемых доходов и количеством совершенных преступлений. Предположительно, чем больше процент сбережений, тем лучше живет население, тем меньше стимулов совершать преступления. Таким образом, предполагаемая зависимость – обратная.
3. Существует зависимость между количеством людей, живущих «за чертой бедности» (с доходом меньше прожиточного минимума) и количеством совершенных преступлений. Это может объясняться тем, что людей, у которых недостаточно денег на нормальное существование, нужда может толкать на преступление. Зависимость – прямая.
4. Существует зависимость между количеством посещений музеев и количеством преступлений. Однако, такая зависимость еще более проявляется в совокупности с количеством выпускников ВУЗов.

## Основные источники данных

Основными источниками данных для нашего исследования являются данных с сайта Генеральной прокуратуры Российской Федерации и с сайта Росстата.

## Публикации

<http://ecsocman.hse.ru/data/493/848/1231/011.DVOIMENNIY.pdf> – Рецидивная преступность: характер, факторы, уровень.

<http://ecsocman.hse.ru/data/058/913/1217/005.SHIPOUNOVA.pdf> – Агрессия и насилие как элементы социальной реальности.

# Предварительный анализ собранных данных

## Анализ особенностей данных

Как отмечалось выше, в качестве параметров мы взяли следующие данные:

* Количество совершенных преступлений;
* Рейтинг качества жизни;
* Численность населения;
* Процент сберегаемых средств;
* Количество безработных людей;
* Численность населения с доходами ниже прожиточного минимума;
* Плотность населения;
* Количество выпускников на ВУЗ на 1000 человек;
* Количество посещений музеев на 1000 человек.

Все данные взяты за 2014 год в разрезе по субъектам федерации. Некоторые данные, такие как **Процент сберегаемых средств**, **Плотность населения** и **Количество выпускников на ВУЗ на 1000 человек** были посчитаны вручную исходя из других данных:

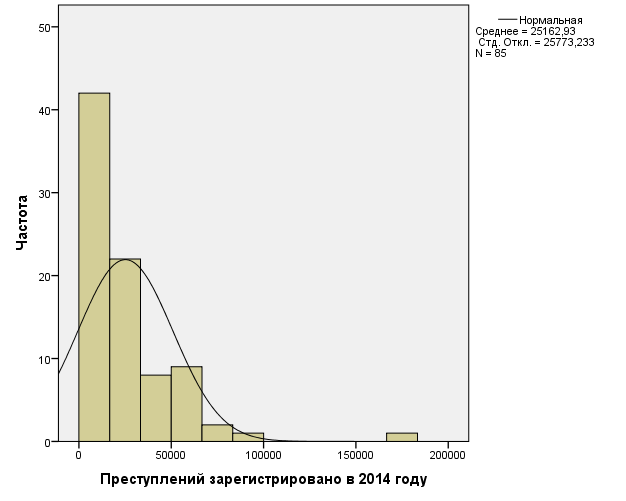
* **Процент сберегаемых средств** = (среднедушевые доходы – среднедушевые расходы) / среднедушевые доходы;
* **Количество выпускников на ВУЗ на 1000 человек** = (Количество выпускников ВУЗов / Количество ВУЗов) / Численность населения \* 1000;
* **Плотность населения** = Численность населения / площадь субъекта федерации.

## Графический анализ данных и описательная статистика

Для большинства переменных, которые задействованы в нашем исследовании подходит графический анализ с использование гистограммы. Кроме того, для количественных переменных будет актуальным посчитать описательные статистики.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Статистика** | | | | | | |
|  | | Количество преступлений | Численность населения в 2014 году (человек) | Процент сбережений от получаемых доходов (%) | Количество безработных (человек) | Количество посещений музеев на 1000 человек в 2014 году | Численность населения с доходами ниже прожиточного минимума (человек) | Количество выпускников на ВУЗ на 1000 человек |
| N | Допустимо | 85 | 85 | 82 | 82 | 84 | 83 | 80 |
| Пропущенные | 0 | 0 | 3 | 3 | 1 | 2 | 5 |
| Среднее значение | | 25162,93 | 1746621,18 | .30336 | 47687,48 | 573,37 | 208,857.841 | 1,529661 |
| Медиана | | 16808,00 | 1191000,00 | .27644 | 34599,50 | 415,00 | 156,496.900 | 1,163257 |
| Стандартная отклонения | | 25773,233 | 1762077,888 | .105700 | 35617,875 | 582,535 | 168,807.3739 | 1,5436061 |
| Асимметрия | | 3,189 | 3,155 | 1,749 | 1,156 | 4,220 | 2,231 | 2,561 |
| Стандартная Ошибка асимметрии | | ,261 | ,261 | ,264 | ,266 | ,263 | ,264 | ,261 |
| Эксцесс | | 15,977 | 14,689 | 3,930 | ,575 | 24,724 | 8,253 | 9,066 |
| Стандартная ошибка эксцесса | | ,517 | ,517 | ,523 | ,526 | ,520 | ,523 | ,517 |
| Минимум | | 645 | 43400 | .129 | 1077 | 61 | 3,906.0 | ,0000 |
| Максимум | | 182873 | 12197600 | .714 | 148755 | 4495 | 1,097,784.0 | 9,4531 |

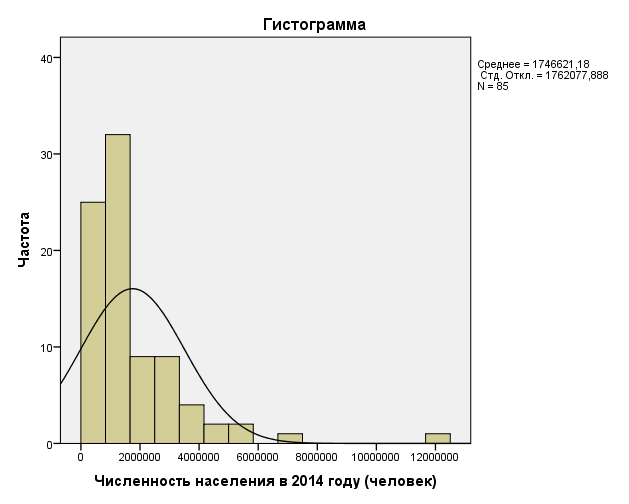
### Количество преступлений



Из гистограммы видим значительную асимметрию влево и островершинность распределения. Из этого, в РФ преобладают регионы, где было совершено достаточно мало преступлений (до 30-35 тысяч за год). Можно заметить, что в данной переменной явно присутствуют выбросы – значение на уровне 175000 (предположительно, Москва) явно отличается от основной массы значений.

Из описательной статистики можно заключить, что данные имеются по всем 85 регионам. Минимальная количество преступлений, зарегистрированных в 2014 году составило 645, а максимальное 182873. Среднее значения равно 25162,93. Медиана же равна 16808,00. Она значительно меньше среднего значение, следовательно, еще раз подтверждается значительная асимметрия влево, что говорит о большем количестве регионов, где было зарегистрировано менее 25 тысяч преступлений. Из-за большой меры разброса, стандартное отклонение переменной также велико. Также имеется значительный положительный эксцесс, что еще раз говорит об островершинности распределения.

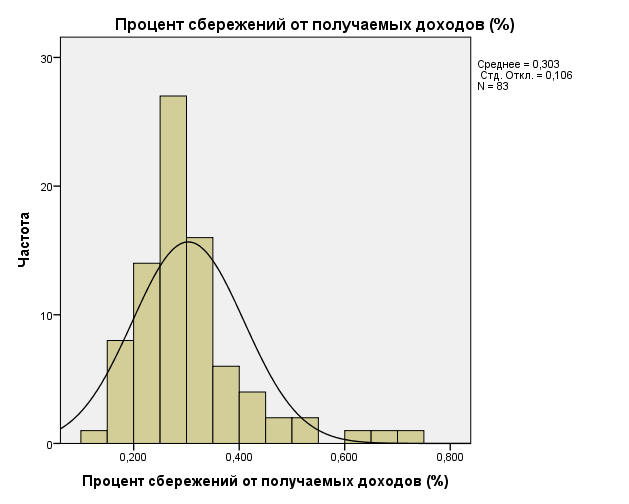
### Численность населения



Из гистограммы видим значительную асимметрию влево и островершинность распределения. Из этого - в РФ преобладают регионы с незначительной населенностью. В данной переменной также присутствуют выбросы – значения на уровне 7 миллионов (предположительно, Санкт-Петербург) и 12 миллионов (предположительно, Москва) явно отличаются от остальных.

Из описательной статистики можно заключить, что данные имеются по всем 85 регионам. Минимальная численность в 2014 году составила 43400 человек, а максимальное 12197600. Среднее значения населенности равно 1746621,18. Медиана же равна 1191000,00. Она меньше среднего значение, следовательно, имеем значительную асимметрию влево, что говорит о большем количестве регионов с численностью населения меньше среднего значения. Из-за большой меры разброса, стандартное отклонение независимой переменной также велико. Также имеется значительный положительный эксцесс, что еще раз говорит об островершинности распределения.

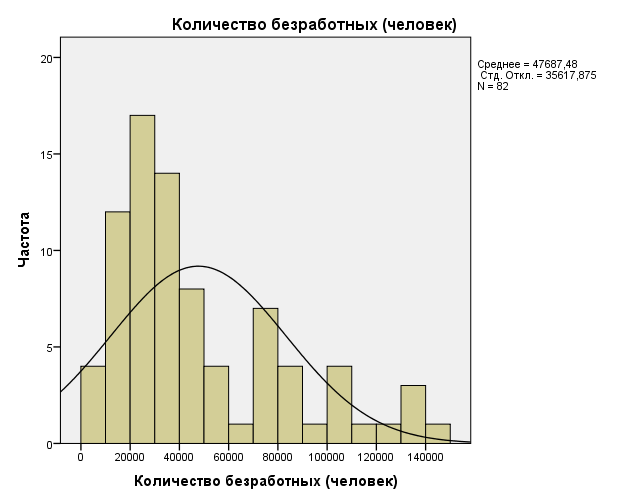
### Процент сбережений от доходов



Из гистограммы видим асимметрию влево и положительный эксцесс, что говорит о преобладании людей с низким процентом сбережений. В отличии от двух предыдущих переменный, в данном случае каких-либо очевидных выбросов не наблюдается. Для более тщательной проверки на предмет выбросов в дальнейшем будут проведены дополнительные тесты.

Из описательной статистики можно заключить, что данные имеются по 83 регионам (пропущено 2 значения). Минимальное значение в 2014 году составило 71,4%, а максимальное 12,9%. Среднее значения равно 30,336%. Медиана же равна 27,644%. Она меньше среднего значение, следовательно, имеем значительную асимметрию влево, что говорит о большем количестве регионов процентом сбережений меньше среднего значения. Стандартное отклонение не слишком велико и равно 10,57%. Также имеется значительный положительный эксцесс, что еще раз говорит об островершинности распределения.

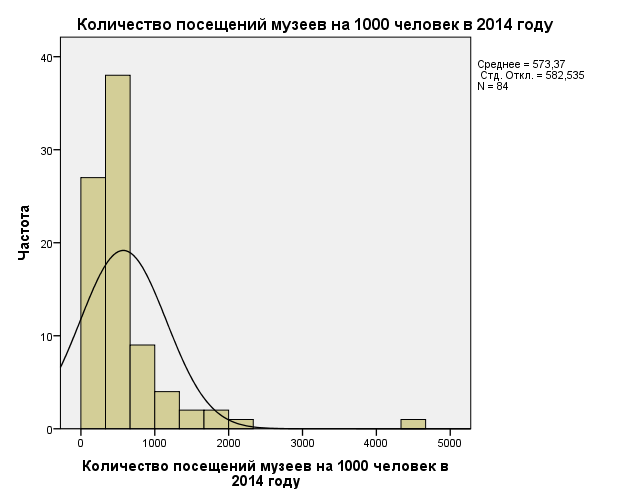
### Количество безработных



Из гистограммы видим незначительную асимметрию влево и положительный эксцесс, что говорит о преобладании регионов с количеством безработных ниже среднего. В данном случае также трудно определить выбросы только по анализу гистограммы. Для этого в дальнейшем првоедем тесты «трех сигм».

Из описательной статистики можно заключить, что данные имеются по 83 регионам (пропущено 2 значения). Минимальное значение в 2014 году составило 1077, а максимальное 148755. Среднее значения равно 47687,48. Медиана же равна 34599,50. Она меньше среднего значение, следовательно, имеем значительную асимметрию влево, что говорит о большем количестве регионов с количеством безработных меньше среднего значения. Стандартное отклонение сравнимо со средним значением и равно 35617,875. Также имеется незначительный положительный эксцесс, что еще раз говорит об островершинности распределения.

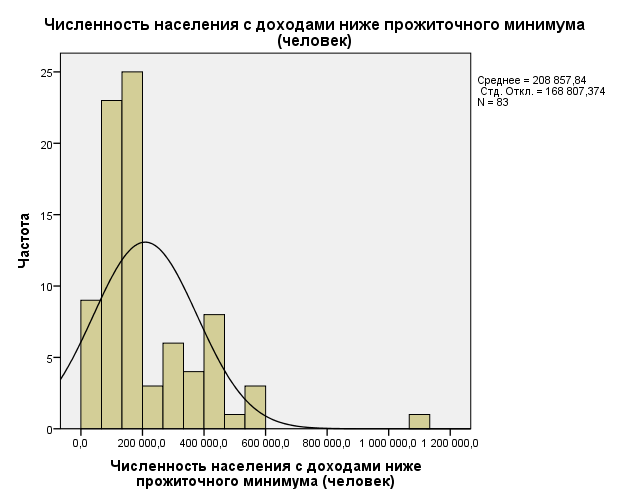
### Количество посещений музеев на 1000 человек



Из гистограммы видим значительную асимметрию вправо и значительный положительный эксцесс. Это говорит о значительном преобладании регионов с низкой посещаемостью музеев. В данном случае наблюдается значительный выброс на уровне 4500 (предположительно, Санкт-Петербург).

Из описательной статистики можно заключить, что данные имеются по 84 регионам (пропущено 1 значение). Минимальное значение в 2014 году составило 61, а максимальное 4495. Среднее значения равно 573,37. Медиана же равна 415,00. Она меньше среднего значение, следовательно, имеем значительную асимметрию влево, что говорит о большем количестве регионов с небольшой посещаемостью музеев. Стандартное отклонение сравнимо со средним значением и равно 582,535. Также имеется значительный положительный эксцесс, что еще раз говорит об островершинности распределения.

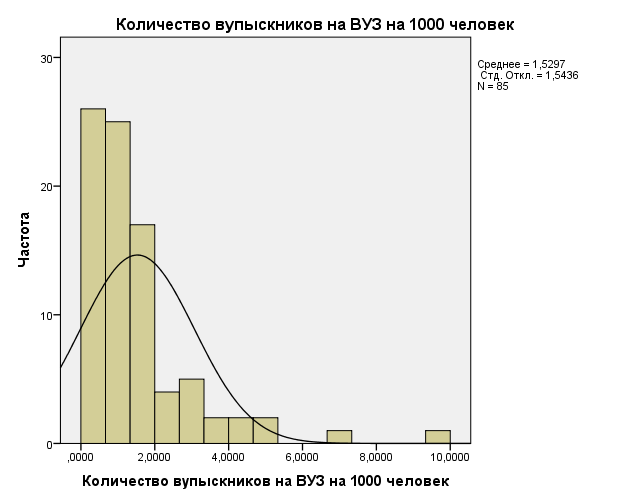
### Численность населения с доходами ниже прожиточного минимума



Из гистограммы видим значительную асимметрию влево и положительный эксцесс, что говорит о большом количестве регионов с уровнем бедности ниже среднего. В данном случае наблюдается значительный выброс на уровне 1 100 000 (предположительно, Москва).

Из описательной статистики можно заключить, что данные имеются по 83 регионам (пропущено 2 значения). Минимальное значение в 2014 году составило 3906, а максимальное 1097784. Среднее значения равно 208857,841. Медиана же равна 156496,900. Она меньше среднего значение, следовательно, имеем значительную асимметрию влево, что говорит о большем количестве регионов с небольшим количеством социально неблагополучных индивидов. Стандартное отклонение сравнимо со средним значением и равно 168807,3739. Также имеется значительный положительный эксцесс, что еще раз говорит об островершинности распределения.

### Количество выпускников на ВУЗ на 1000 человек



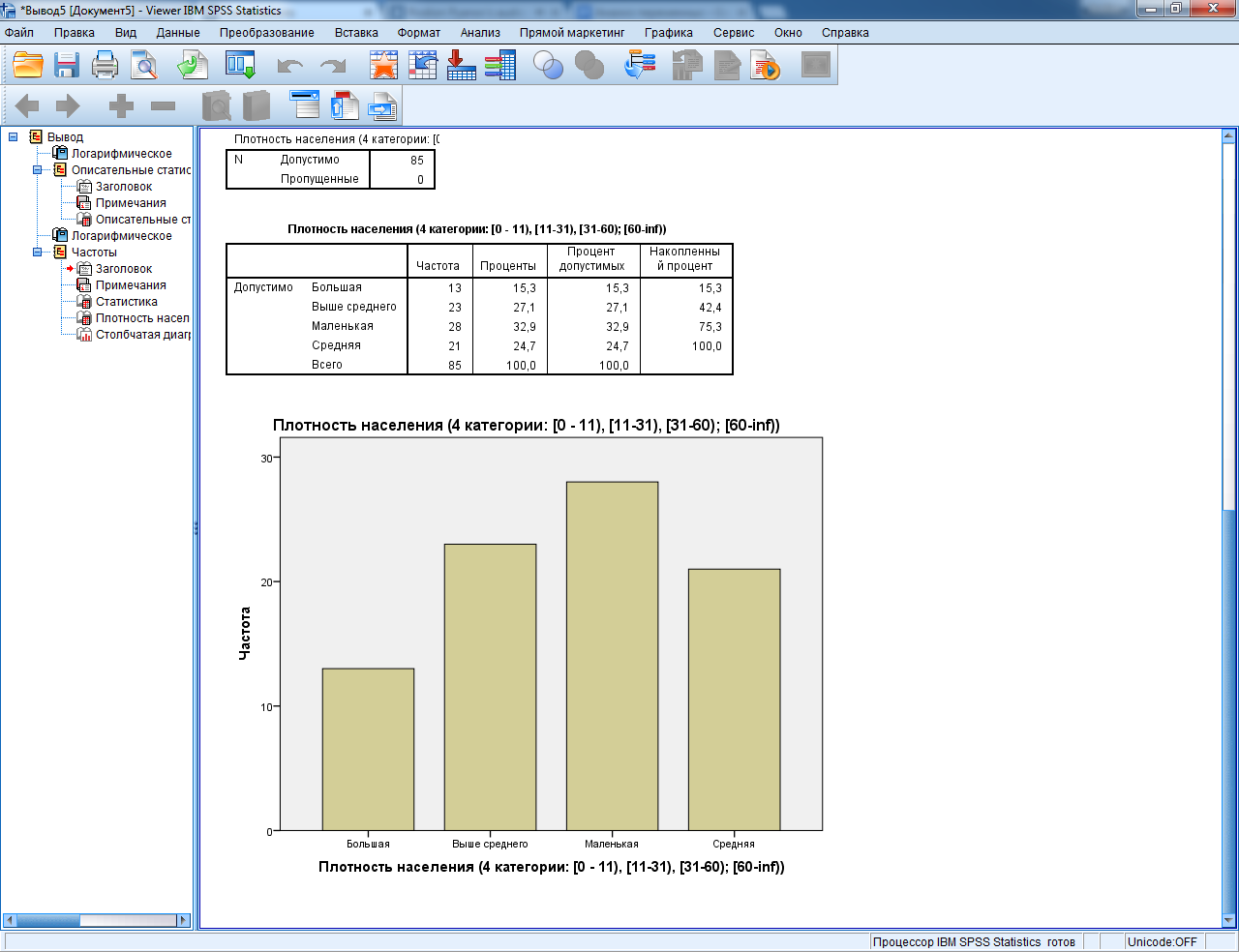
Из гистограммы видим асимметрию влево и положительный эксцесс, что говорит, о преобладании регионов с небольшим количеством высококвалифицированных кадров. В данном случае наблюдается два довольно значительных выброса на уровне 7 (предположительно, Магаданская область) и на уровне 9,5 (предположительно, Республика Калмыкия).

Из описательной статистики можно заключить, что данные имеются по всем 80 регионам (пропущено 5 значений). Минимальное значение в 2014 году составило 0, а максимальное 9,4531. Среднее значения равно 1,529661. Медиана же равна 1,163257. Она меньше среднего значение, следовательно, имеем значительную асимметрию влево, что говорит о большем количестве регионов с небольшим количеством высокообразованных индивидов. Стандартное отклонение сравнимо со средним значением и равно 1,5436061. Также имеется значительный положительный эксцесс, что еще раз говорит об островершинности распределения.

### Плотность населения

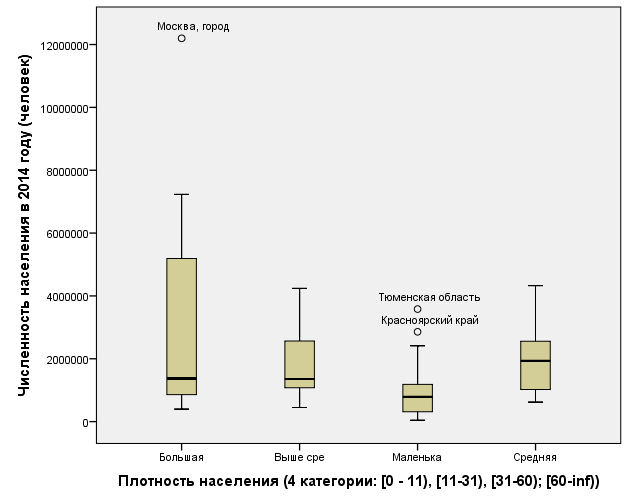
Для **Плотности населения** проведем графический анализ с помощью гистограммы, а также при помощи диаграмм Бокса-Вискера, на которых рассмотрим данную переменную в сочетании с другими. Это сделано для того, чтобы наглядно увидеть и проанализировать причины возникающих выбросов (однако, кроме этого, далее будет проведен анализ выбросов по методу трех сигм) и детальнее понять исходные данные.

Кроме этого стоит упомянуть о том, как была построена данная переменная. Формула Стерджесса предписывала разделить переменную на 7 или 8 групп, однако, подобное разделение показалось нам избыточным. В итоге, мы решили разделить **Плотность населения** на 4 группы. Сами группы сформировали по следующему признаку (Маленькая плотность – [0;11) человек на квадратный километр, Средняя плотность – [11;31) человек на квадратный километр, Плотность выше среднего – [31;60) человек на квадратный километр, Большая плотность – больше 60 человек на квадратный километр.



Из гистограммы выше видим асимметрию вправо. При этом преобладают регионы с плотностью населения от 0 до 11 человек на квадратный километр.

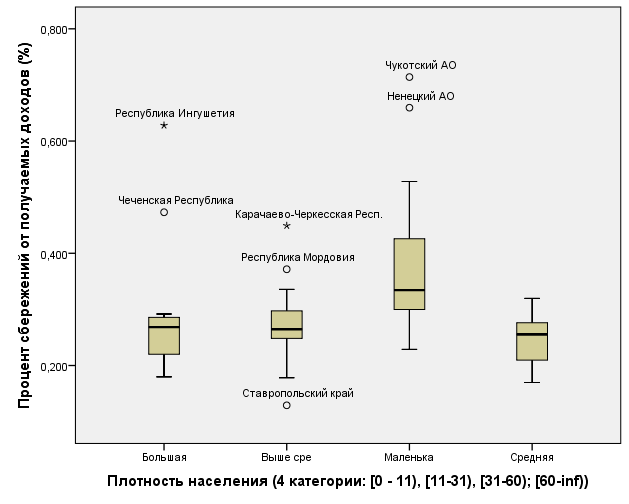
#### Плотность населения и Численность населения



Одним из наиболее сильных выбросов является Москва. Это объясняется тем, что город является столицей и в нем сконцентрировано почти все рабочие места. Следовательно, плотность населения здесь намного выше чем в любом другом городе РФ. Почти все, кто желает заработать денег, тянутся в столицу и, при изменениях в экономической ситуации, люди все больше поддаются данной тенденции. Остальные выбросы объясняются лишь размерами территорий, отведенных данным регионам и малого числа крупных населенных пунктов, в которых сконцентрировано производство и, соответственно, население.

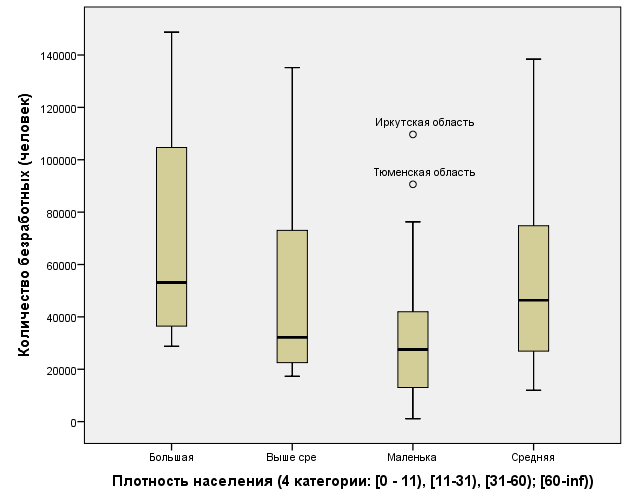
В данном случае группы «Большая» и «Выше среднего» можно было бы объединить в одну, т.к. их медианы почти не различаются, а «Выше среднего» полностью входит по значениям в границы «Большой». Однако, мы посчитали нецелесообразным выделять только три группы вместо четырех, т.к. на при построении подобной диаграммы для других переменных было обнаружено, что «Большая» плотность и плотность «Выше среднего» довольно сильно отличаются по своим характеристикам.

#### Плотность населения и Процент сбережений от полученных доходов



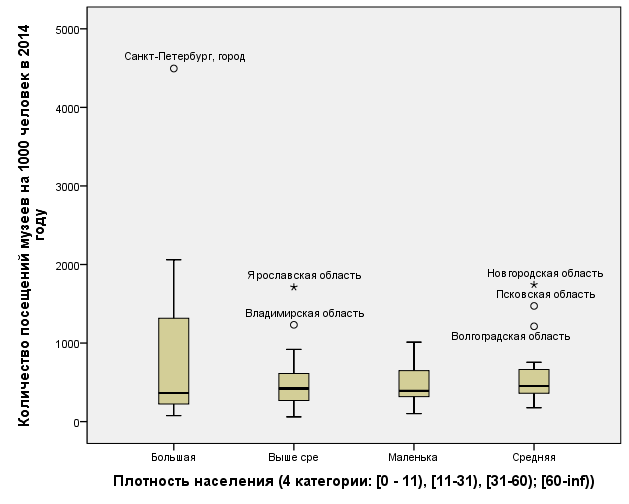
Причиной показанных выбросов, вероятнее всего, является тот факт, что в данных регионах не сильно развита денежная система. В них люди выживают и используют ресурсы лишь для выживания. В некоторых других регионах люди иначе распоряжаются своими средствами, возможно, в виду особенностей рынка в данной области или внутренних культурных традиций. В данном случае ни одну из групп нельзя объединить с другой из-за сильного различия в максимальных и минимальных значениях между группами, а также из-за разницы медиан в этих группах.

#### Плотность населения и Количество безработных



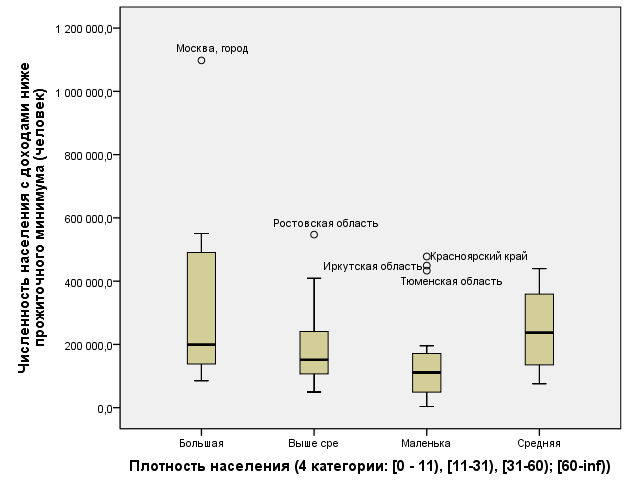
Причиной выбросов на данной диаграмме могут быть особые обстоятельства внутри данных регионах в исследуемом периоде времени. Найти более подробную информацию, объясняющие данные выбросы нам, к сожалению, не удалось. В данном случае ни одну из групп нельзя объединить с другой из-за сильного различия в максимальных и минимальных значениях между группами, а также из-за разницы медиан в этих группах.

#### Плотность населения и Количество посещений музеев на 1000 человек



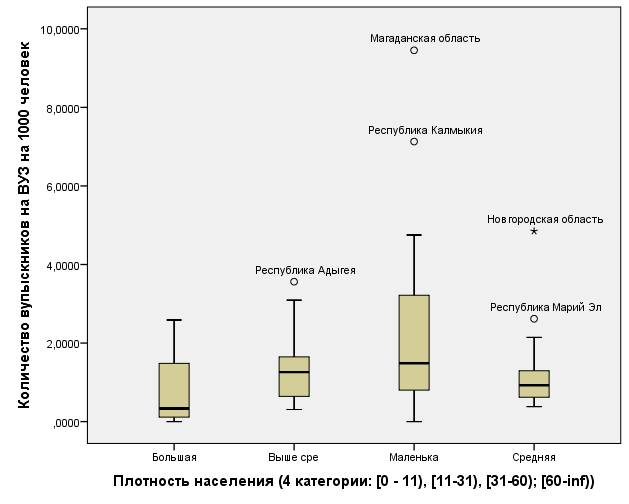
На данной диаграмме основными выбросами являются области с самым большим культурным наследством нашей Родины. В них сохранились многие памятники старины, к которым съезжаются туристы и за счет этого число посещений музеев данных областей выходят за рамки общей тенденции. Также стоит отметить самый большой выброс в культурной столице страны, где находятся самые богатые музеи и экспонаты, которые привлекают тысячи туристов с разных концов света. В данном случае, в принципе, можно было бы объединить все группы в одну, т.к. медианы у всех групп находятся примерно на одном уровне, а по максимальным и минимальным значениям они все входят в границы «Большой» группы. Однако, также, как и в ситуации с «Численностью населения», мы решили, что такое объединение не принесет пользы из-за значительного различия групп в диаграммах Бокса-Вискера по другим переменным.

#### Плотность населения и Численность населения с доходом ниже прожиточного минимума



Многие крупные города содержат большое количество людей поэтому и численность безработных в них выше, поэтому выделяющиеся значения в диаграмме численности населения совпадают с данной диаграммой. Однако некоторые города так же попали в выброс в виду крайне плачевной ситуацией в производстве и общем числом рабочих мест в области. В данном случае ни одну из групп нельзя объединить с другой из-за сильного различия в максимальных и минимальных значениях между группами, а также из-за разницы медиан в этих группах.

#### Плотность населения и Количество выпускников на ВУЗ на 1000 человек

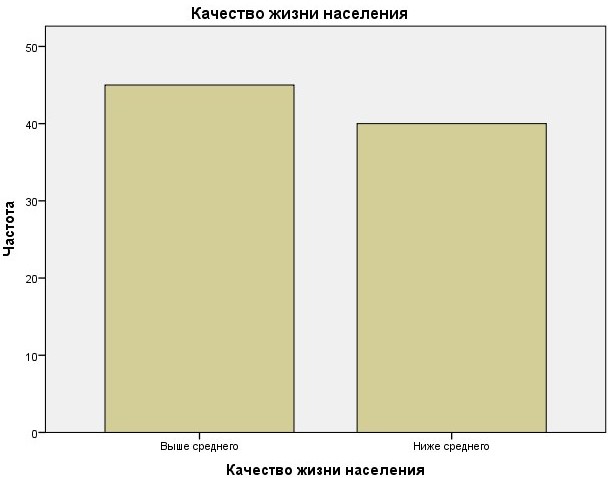


Данные выбросы могут быть связаны с тем, что в показанных субъектах федерации или крайне мало высших учебных заведений, или строгость оценки и компетентность ВУЗов находится на низком уровне. Как следствие, малое количество вузов выпускает большое количество людей с высшим образованием. В данном случае ни одну из групп нельзя объединить с другой из-за сильного различия в максимальных и минимальных значениях между группами, а также из-за разницы медиан в этих группах.

### Качество жизни населения

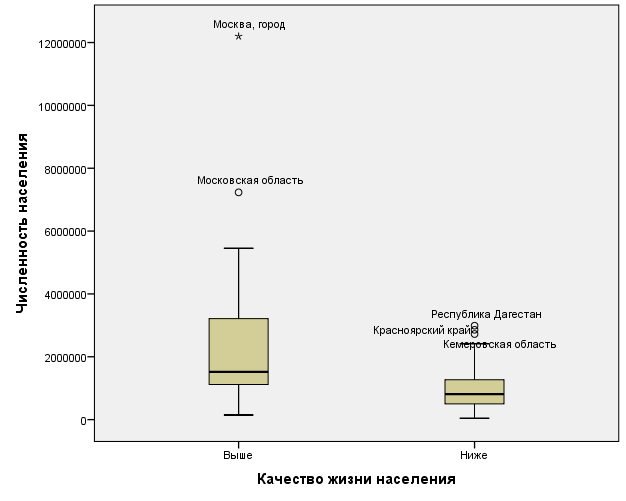
Для **Качества жизни населения** проведем графический анализ с помощью гистограммы, а также при помощи диаграмм Бокса-Вискера, на которых рассмотрим данную переменную в сочетании с другими. Это сделано для того, чтобы наглядно увидеть и проанализировать причины возникающих выбросов (однако, кроме этого, далее будет проведен анализ выбросов по методу трех сигм) и детальнее понять исходные данные.

Кроме этого стоит упомянуть о том, как была построена данная переменная. Формула Стерджесса предписывала разделить переменную на 7 или 8 групп, однако, подобное разделение показалось нам избыточным. В итоге, мы решили разделить **Качество жизни населения** всего на 2 группы – «Выше среднего» и «Ниже среднего». Сами данные первоначально представляли рейтинг, где каждому субъекту федерации был сопоставлен определенный балл, полученный из 60-ти других значений и их комбинаций. Разделение на вышеуказанные группы происходило по очевидному сценарию – субъекты, у которых рейтинг был выше среднего значения рейтинга всех субъектов, попали в первую группу, остальные – во вторую.



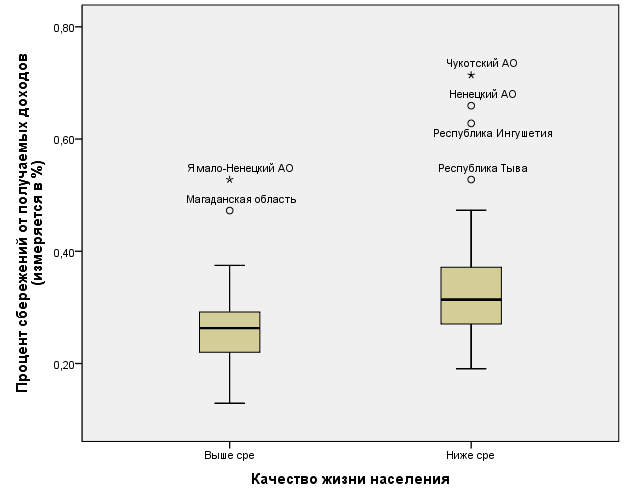
Из гистограммы выше видим незначительную асимметрию влево. При этом преобладают регионы с качеством жизни «Выше среднего».

#### Качество жизни населения и Численность населения



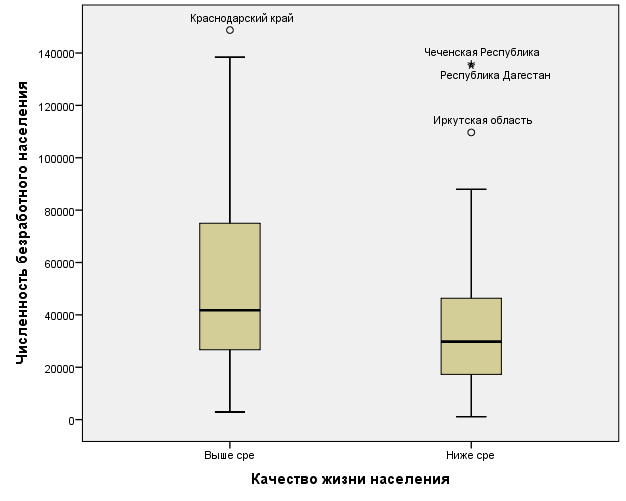
Одними из наиболее сильных выбросов являются Москва и Московская область. Это объясняется тем, что в Москве и МО сконцентрировано большинство рабочих мест. Следовательно, численность населения здесь гораздо больше, чем в любом другом субъекте РФ. Остальные выбросы объясняются лишь размерами территорий, отведенных данным регионам и малого числа крупных населенных пунктов, в которых сконцентрировано производство и, соответственно, население. Данные группы нецелесообразно объединять в одну из-за различающихся значений медиан внутри этих групп.

#### Качество жизни населения и Процент сбережений от полученных доходов



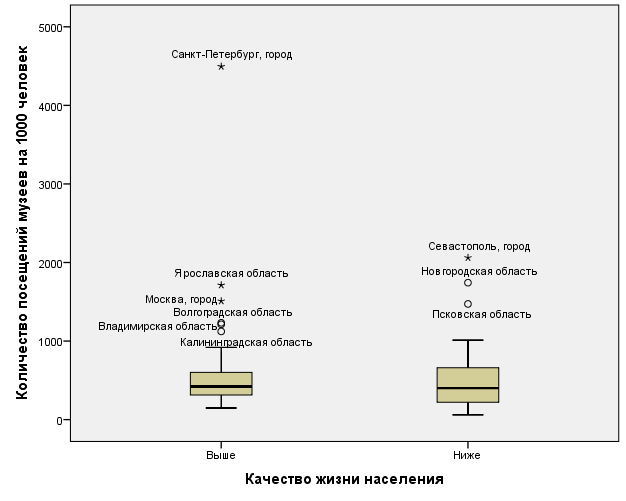
Причиной показанных выбросов, вероятнее всего, является тот факт, что в данных регионах не сильно развита денежная система. В них люди выживают и используют ресурсы лишь для выживания. В некоторых других регионах люди иначе распоряжаются своими средствами, возможно, в виду особенностей рынка в данной области или внутренних культурных традиций. В данном случае объединять две группы нет смысла все по той же причине – слишком большое отличие медиан внутри групп.

#### Качество жизни населения и Количество безработных



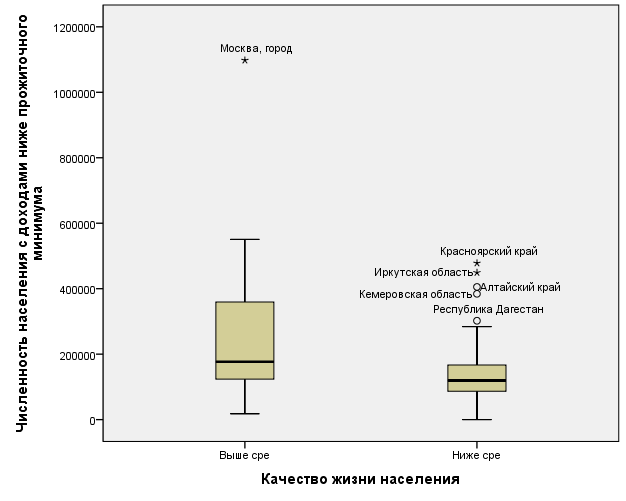
Причиной выбросов на данной диаграмме могут быть особые обстоятельства внутри данных регионах в исследуемом периоде времени. Например, в Краснодарском крае может быть большое число граждан, которые зарабатывают, сдавая жилье отдыхающим во время летнего сезона и, таким образом, не нуждаются в другом источнике дохода. С Чеченской Республикой и Республикой Дагестан дела могут обстоять по-другому – люди, которые хотят работать и развиваться, как правило, уезжают в другие регионы (например, в Москву), оставляя на своей малой родине своих не столь амбициозных товарищей. Объединение групп в данном случае все также нецелесообразно, по той же причине – большому различию медиан внутри групп.

#### Качество жизни населения и Количество посещений музеев на 1000 человек



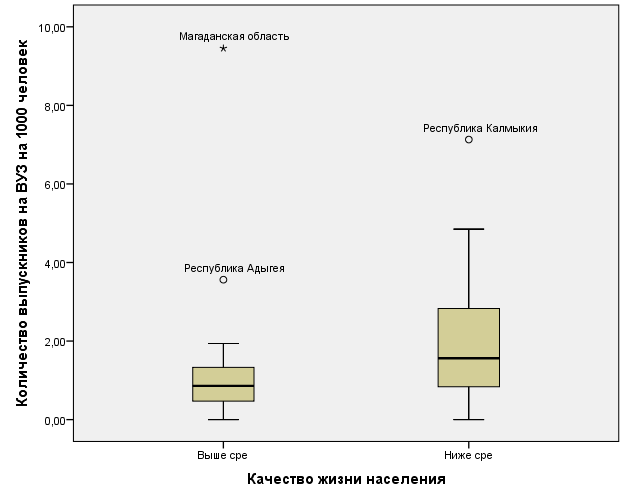
На данной диаграмме основными выбросами являются области с самым большим культурным наследством нашей Родины. В них сохранились многие памятники старины, к которым съезжаются туристы и за счет этого число посещений музеев данных областей выходят за рамки общей тенденции. Также стоит отметить самый большой выброс в культурной столице страны, где находятся самые богатые музеи и экспонаты, которые привлекают тысячи туристов с разных концов света. В данном случае, в принципе, можно было бы объединить две группы в одну, т.к. медианы у них находятся примерно на одном уровне, а по максимальному и минимальному значениям «Выше среднего» полностью входит в границы «Ниже среднего». Однако, понимая, что объединение в данной случае приведет к объединению во всех остальных, мы не стали этого делать, т.к. в других группах объединение недопустимо.

#### Качество жизни населения и Численность населения с доходом ниже прожиточного минимума



Москва, как самый крупный город в России, содержит большое количество людей поэтому, очевидно, и численность безработных в Москве выше. Однако некоторые города так же попали в выброс в виду крайне плачевной ситуацией в производстве и общем числе рабочих мест в области. В данном случае группы нельзя объединить из-за сильного различия в максимальных и минимальных значениях между группами, а также из-за разницы медиан в этих группах.

#### Качество жизни населения и Количество выпускников на ВУЗ на 1000 человек



Данные выбросы могут быть связаны с тем, что в показанных субъектах федерации или крайне мало высших учебных заведений, или строгость оценки и компетентность ВУЗов находится на низком уровне. Как следствие, малое количество вузов выпускает большое количество людей с высшим образованием. В данном случае ни одну из групп нельзя объединить с другой из-за сильного различия в максимальных и минимальных значениях между группами, а также из-за разницы медиан в этих группах.

## Пропущенные значения и выбросы

Исходя из данных, полученных из описательной статистики, мы можем сделать вывод, что 5 из 85 значений являются пропущенными хотя бы одной из переменных. Следовательно, исключаем их дальнейшего исследования, чтобы отсутствие данных не помешало нам в дальнейшем. Субъекты федерации, у которых есть пропущенные значения: Ненецкий АО, Республика Крым, город Севастополь, Чукотский АО, Ямало-Ненецкий АО.

Что касается выбросов, то используя метод трех сигм (данные для него были получены на этапе построения гистограмм), находим следующие выбросы:

* По **Количеству преступлений** – город Москва
* По **Численности населения** – город Москва и Московская область
* По **Проценту сбережений от доходов** – Республика Ингушетия
* По **Количеству безработных** – выбросов нет
* По **Количеству посещений музеев на 1000 человек** – город Санкт-Петербург
* По **Численности населения с доходами ниже прожиточного минимума** – город Москва
* По **Количеству выпускников на ВУЗ на 1000 человек** – Магаданская область, Республика Калмыкия

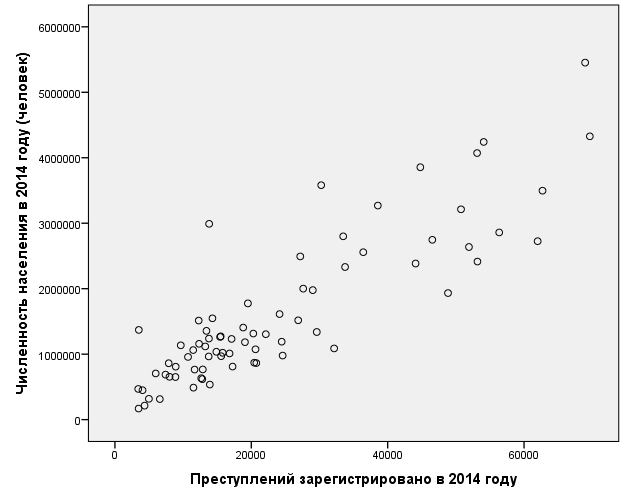
Следовательно, как и пропущенные значения, исключим из дальнейшего исследования и выбросы.

## Анализ статистической связи

### Графический анализ

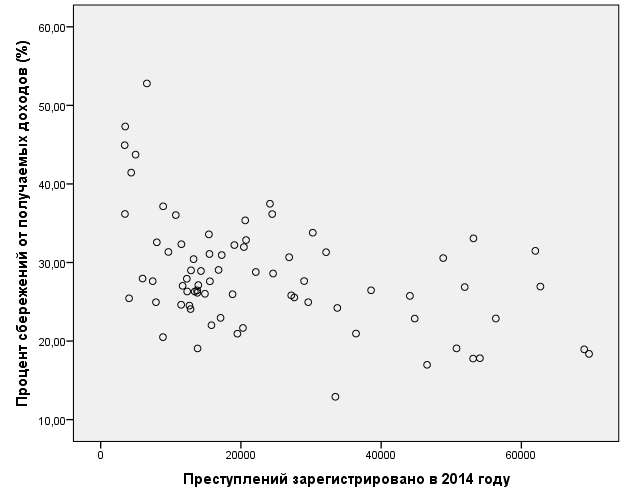
Теперь рассмотрим наличие связи и её характер для зависимой переменной и остальных.

#### Численность населения



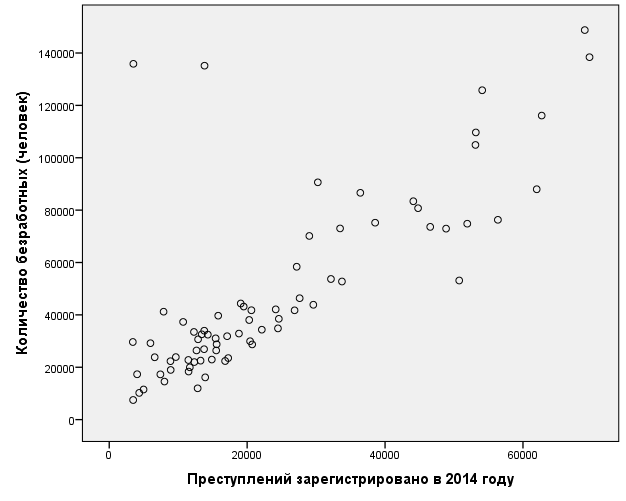
Очевидно, что численность населения напрямую влияет на количество преступлений. Этот простой и логичный вывод подтверждает соответствующая диаграмма рассеивания. В данном случае зависимость прямая и, вероятнее всего, линейная.

#### Процент сбережений от получаемых доходов



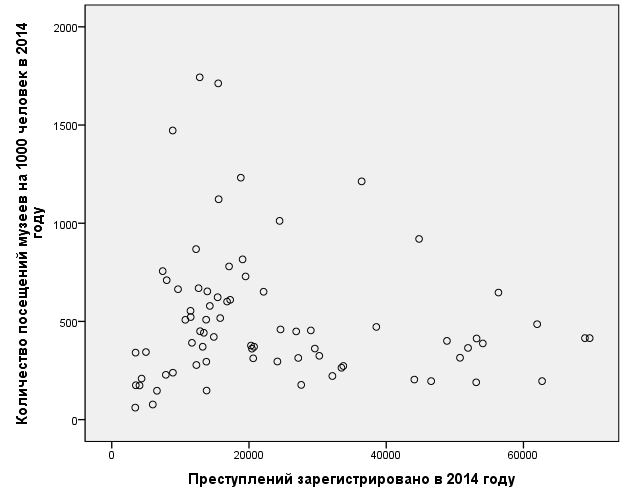
В данном случае диаграмма рассеивания показывает, что зависимость обратная. Это логично, ведь чем меньше человек сберегает, тем больше он тратит. Следовательно, ему может не хватать денег и это может толкнуть его на преступление. По своей форме зависимость напоминает гиперболическую или линейную.

#### Количество безработных



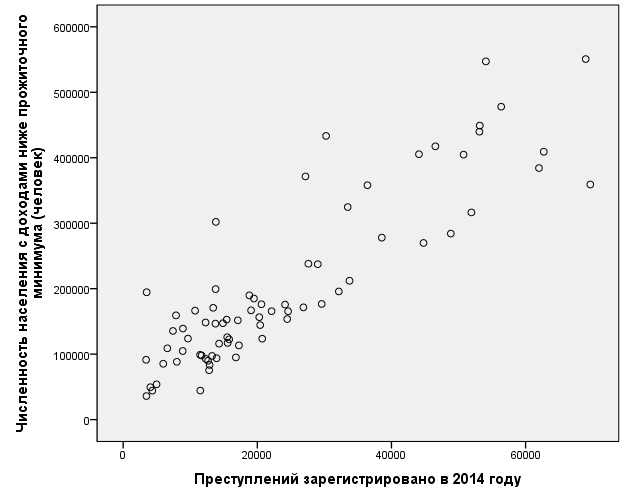
В случае с количеством безработных зависимость прямая. Чем больше безработных – тем больше преступлений, что тоже вполне логично. У безработных нет путей получения денежных средств, кроме как незаконными путями. По своей форме зависимость похожа на линейную или степенную (корень второй степени).

#### Количество посещений музеев



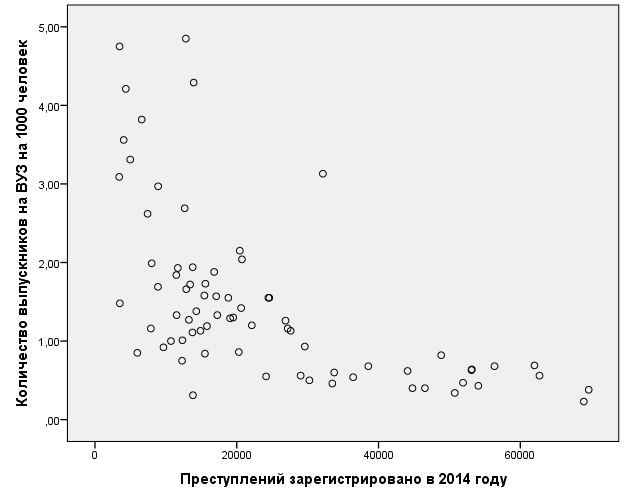
В данном случае зависимость неочевидна. С одной стороны, на данной диаграмме рассеивания можно заметить прямую линейную или степенную (корень второй степени) зависимость. Однако, прямая зависимость, какой бы формы она не была, не поддается логическому объяснению и, скорее, контринтуитивна.

#### Численность населения с дохода ниже прожиточного минимума



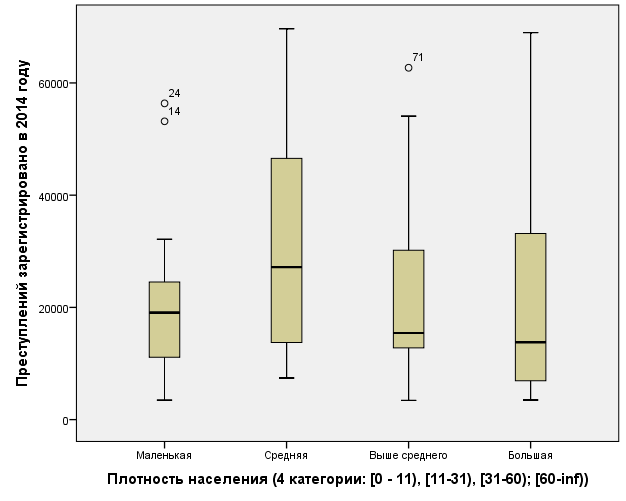
В данном случае наблюдается прямая линейная зависимость, хорошо поддающаяся объяснению –людям не хватает имеющихся у них денежных средств на нормальное существование и это приводит их к получению дохода путем совершения преступлений (разбои, грабежи и т.д.).

#### Количество выпускников на ВУЗ на 1000 человек



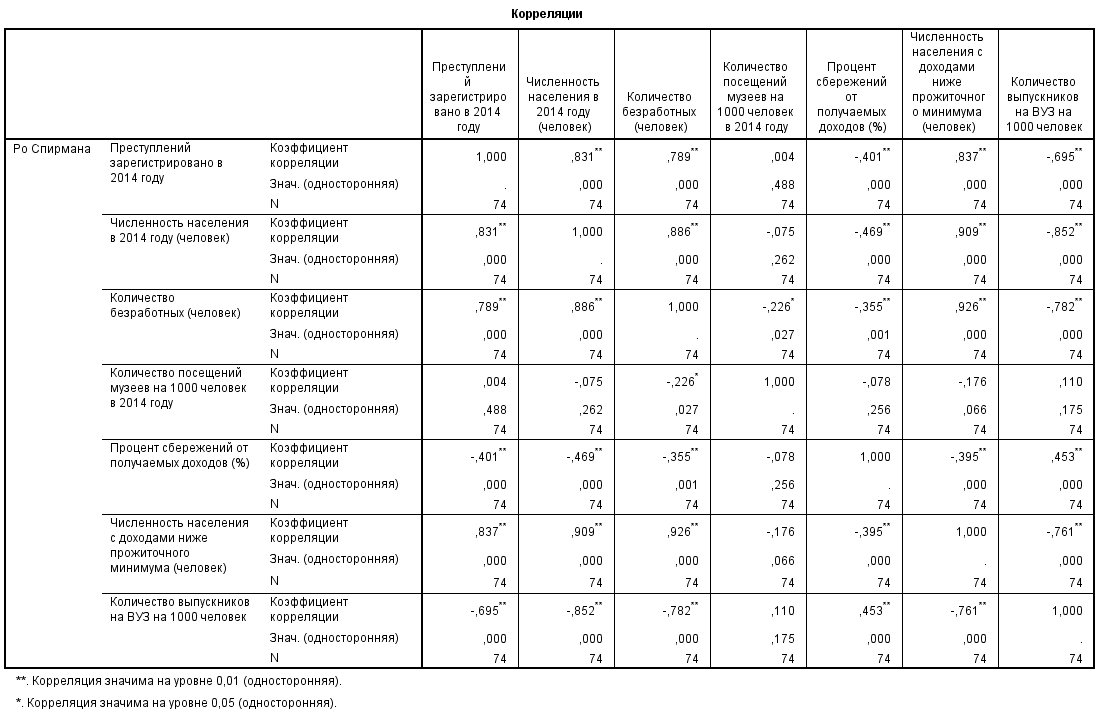
В данном случае наблюдается обратная, вероятнее всего, гиперболическая зависимость. Эта зависимость тоже поддается объяснению – чем больше человек с высшим образованием выпускает ВУЗ, тем больше людей смогут найти высокооплачиваемую работу и не будут нуждаться в доходе, получаемым путем совершения противоправных действий.

#### Плотность населения



В данном случае видно, что наибольшее количество преступлений совершается в городах со средней плотностью населения. Это можно объяснить тем, что в городах с более высокой плотностью населения, как правило, больше сил правопорядка, которые затрудняют совершение преступления. В городах же с маленькой плотностью преступлений может быть меньше из-за банальной трудности в поиске жертвы.

### Корреляционная матрица



Объясним корреляцию Количества преступлений с остальными числовыми переменными.

#### Количество преступлений и Численность населения

В данном случае присутствует значительная корреляция с коэффициентом по Спирману = 0,831. Данная корреляция значима на уровне 0,01 – т.е., гипотеза о нулевой корреляции отвергается. Также стоит отметить, что данная корреляция прямая – т.е., значение Количества преступлений меняется в ту же сторону, что и Численность населения.

#### Количество преступлений и Количество безработных

В данном случае присутствует значительная корреляция с коэффициентом по Спирману = 0,789. Данная корреляция значима на уровне 0,01 – т.е., гипотеза о нулевой корреляции отвергается. Также стоит отметить, что данная корреляция прямая – т.е., значение Количества преступлений меняется в ту же сторону, что и Количество безработных.

#### Количество преступлений и Количество посещений музеев

В данном случае корреляция почти отсутствует. Коэффициент по Спирману = 0,004. Однако стоит отметить, что данная корреляция не значима на приемлемых уровнях и, следовательно, гипотеза о нулевой корреляции НЕ отвергается.

#### Количество преступлений и Процент сбережений

В данном случае присутствует слабая корреляция с коэффициентом по Спирману = -0,401. Данная корреляция значима на уровне 0,01 – т.е., гипотеза о нулевой корреляции отвергается. Также стоит отметить, что данная корреляция обратная – т.е., значение Количества преступлений меняется в обратную сторону относительно Процента сбережений.

#### Количество преступлений и Численность населения с доходами ниже прожиточного минимума

В данном случае присутствует значительная корреляция с коэффициентом по Спирману = 0,837. Данная корреляция значима на уровне 0,01 – т.е., гипотеза о нулевой корреляции отвергается. Также стоит отметить, что данная корреляция прямая – т.е., значение Количества преступлений меняется в ту же сторону, что и Численность населения с доходами ниже прожиточного минимума.

#### Количество преступлений и Количество выпускников на ВУЗ на 1000 человек

В данном случае присутствует сильная корреляция с коэффициентом по Спирману = -0,695. Данная корреляция значима на уровне 0,01 – т.е., гипотеза о нулевой корреляции отвергается. Также стоит отметить, что данная корреляция обратная – т.е., значение Количества преступлений меняется в обратную сторону относительно Количества выпускников на ВУЗ на 1000 человек.

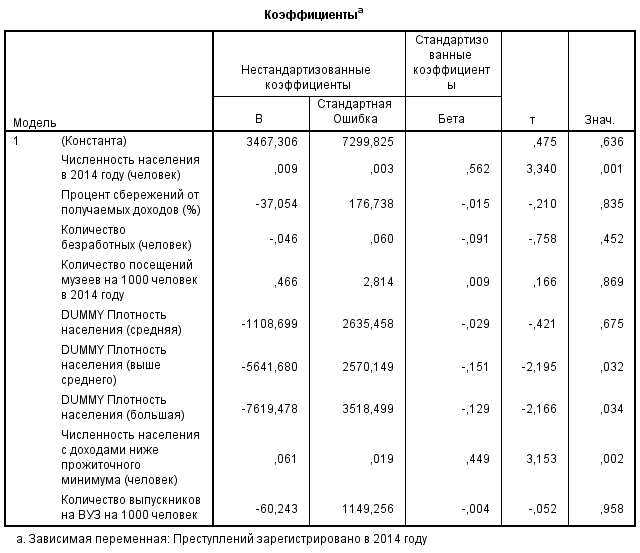
## Вывод из предварительного анализа данных

Проведя предварительный анализ данных, мы смогли обнаружить и исключить пропущенные значения и выбросы, а также более детально понять наши данные. Кроме того, мы подкрепили наши первоначальные гипотезы различными исследованиями и можем убедиться, что первые 3 гипотезы действительно имеют право на существование. Что касается четвертой гипотезы, к сожалению, она не нашла подтверждения в ходе нашего исследования.

# Построение модели

## 3.1. Базовая модель

Следующим шагом после проведения предварительного анализа было построение линейной регрессии. Стоит отметить, что перед тем, как непосредственно построить регрессии в SPSS, мы превратили зависимые и независимые номинальные переменные в «дамми» переменные, создав для каждого значения номинальной переменной отдельный столбец (переменную) и заполнили новые столбцы нулями и единицами в соответствии с наличием или отсутствием фактора.



### 3.1.1. Связь структуры базовой модели с основными гипотезами исследования

Напомним основные гипотезы, выдвинутые ранее и трансформируем их в статистические:

1. Первая, самая очевидная гипотеза заключается в наличии прямой зависимости между количеством совершенных преступлений и количеством населения в конкретном субъекте федерации => **переменная «Численность население в 2014 году» значима в модели и ее коэффициент положителен**
2. Существует зависимость между процентом сбережений от получаемых доходов и количеством совершенных преступлений. Предположительно, чем больше процент сбережений, тем лучше живет население, тем меньше стимулов совершать преступления => **переменная «Процент сбережений от получаемых доходов» значима в модели и ее коэффициент отрицателен**
3. Существует зависимость между количеством людей, живущих «за чертой бедности» (с доходом меньше прожиточного минимума) и количеством совершенных преступлений. Это может объясняться тем, что людей, у которых недостаточно денег на нормальное существование, нужда может толкать на преступление => **переменная «Численность населения с доходами ниже прожиточного минимума» значима в модели и ее коэффициент положителен**
4. Существует зависимость между количеством посещений музеев и количеством преступлений. Однако, такая зависимость еще более проявляется в совокупности с количеством выпускников ВУЗов => **коэффициент корреляции между переменными «Количество посещений музеев» и «Количество выпускников вузов» близок к единице и переменная «Количество посещений музеев» значима в модели**

### 3.1.2. Связь структуры базовой модели с результатами предварительного анализа данных

Во время проведения предварительного анализа данных были определены функциональная форма вхождения независимых переменных в итоговую модель. В ходе построения линейной регрессии были получены действительные коэффициенты и уровень значимости каждой переменной. Таким образом, существует возможность сравнения структуры базовой модели с результатами предварительного анализа данных.

По результатам предварительного анализа были получены следующие данные о связи «Количества преступлений» с остальными переменными:

* С «Численностью населения в 2014 году» - прямая, линейная зависимость
* С «Процентом сбережений от получаемых доходов» - обратная, нелинейная зависимость
* С «Количеством безработных» - прямая, линейная зависимость
* С «Количеством посещений музеев на 100 человек» - нет зависимости
* С «Численностью населения с доходами ниже прожиточного минимума» - прямая, линейная зависимость
* С «количеством выпускников на ВУЗ на 100 человек» - обратная, нелинейная зависимость

По результатам корреляционной таблицы и базовой регрессии можно судить, что все предположения, сделанные на этапе предварительного анализа данных, оказались правильными, кроме предположения о наличии линейной зависимости «Количества преступлений» и «Количества безработных» - регрессия отвергла существование линейной зависимости.

### 3.1.3. Оценивание базовой модели

С переменными, которые входят в регрессию, мы провели ряд тестов на гетероскедастичность и мультиколлинеарность.

#### 3.1.3.1. VIF

Первый был VIF-тест на мультиколлинеарность. С помощью SPSS были построены таблицы следующего вида для всех переменных:



Приводить еще 8 подобных таблиц в нашем отчете мы не будем, лишь отметим, что во всех таблицах ни одного значения VIF, больше чем 10 не наблюдалось, что свидетельствует об отсутствии (или малой степени) мультиколлинеарности.

#### 3.1.3.2. Индекс обусловленности

Вторым способом проверить мультиколлинеарность является индекс обусловленности. Для него необходимо взять матрицу, столбцами которой являются задействованные в регрессии переменные, а строчками – значения этих переменных. Затем матрицу необходимо транспонировать, а потом – умножить транспонированную матрицу на обычную. После выполнения всего описанного выше, мы получили следующую матрицу:



Далее, для нахождения индекса неопределённости необходимо найти собственные значения этой матрицы и разделить корень из максимального из них на корень из минимального из них.

В нашем случае это значение равно 120,82. Оно значительно больше 30, поэтому по индексу обусловленности в данных присутствует мультиколлинеарность.

Исходя из двух тестов, которые дали противоположный результат, а также из того, что получившийся индекс обусловленности значительно выше порогового значения, было принято решение о наличии мультиколлинеарности в данной модели.

#### 3.1.3.3. Тест Парка

В данном тесте на гетероскедастичность сначала необходимо построить обычную регрессию и понять коэффициенты каждой переменной. Затем нужно вычислить по этим коэффициентам предсказанное значение и сравнить его с наблюденным. Таким образом, мы получим ошибку, которую следует возвести в квадрат. Когда все это сделано, остается лишь взять логарифм от ошибки в квадрате и всех остальных независимых переменных и построить парные регрессии (логарифм от каждой независимой переменной с логарифмом ошибки в квадрате). Определяющее значение здесь будет иметь значимость логарифма от независимой переменной в каждой регрессии. Если логарифм от независимой переменной значим, то гетероскедастичность есть.

В данном тесте, проведенном на наших данных, значения p-value во трех регрессиях оказались больше, чем 0,05 и в трех – значительно больше, чем 0,05. Это говорит о том, что тест Парка не может дать однозначный ответ относительно наличия или отсутствия гетероскедастичности в наших данных. Соответственно, необходимо провести другие тесты.

#### 3.1.3.4. Тест Breusch-Pagan

Данный тест также помогает выявить наличие гетероскедастичности в исследуемых данных. Для проведения теста необходимо, как и в тесте Парка, найти квадрат ошибок. Затем нужно построить регрессию, где целевой переменной будет не «Количество преступлений», а найденный квадрат ошибки. Затем, следует умножить значение R-квадрат, полученной из регрессии на количество наблюдений и проверить полученное значение на распределение ХИ-квадрат с количеством степеней свободы, равным количеству независимых переменных. Если значение распределено как ХИ-квадрат – гетероскедастичности нет.

В нашем случае при проверке значения на распределение ХИ-квадрат, мы получили 0,02, что меньше необходимых 0,05. Это говорит о том, что значение не распределено как ХИ-квадрат, а значит – гетероскедастичность есть.

#### 3.1.3.5. Тест Goldfeld-Quandt

Данный тест является еще одним способом выявить наличие гетероскедастичности в нашей выборке. Для проведения этого теста необходимо выбрать одну из зависимых переменных и отсортировать всю выборку по возрастанию данной переменной. Затем, следует разделить выборку на 3 части и удалить среднюю часть. Таким образом, остается две части выборки – с минимальными значениями выбранной зависимой переменной и максимальными. Далее необходимо построить обычные регрессии для обоих частей и найти их RSS. После этого остается лишь поделить RSS2 (которая, как правило, больше, чем RSS1) на RSS1 и проверить полученной значение на F-распределение. Если F-распределение подтверждается, то гетероскедастичности нет.

В нашем случае мы проделали данный тест три раза, используя различный независимые переменные. Во всех случаях полученное значение не было распределено в соответствии с F-распределением, что говорит о наличии гетероскедастичности.

## 3.1.2 Проверка гипотез для базовой модели

1) Первая гипотеза значима на уровне значимости 1% (следовательно, и на уровне 5% и 10%), т.к. значимость переменной «Численность населения» находится на уровне 0,001 и ее коэффициент равняется 0,009, что больше нуля.

2) Вторая гипотеза отвергается, т.к. значимость переменной «Процент сбережений от получаемых доходов» равняется 0,835.

3) Третья гипотеза, также, как и первая, принимается на уровне значимости 1% (следовательно, и на уровне 5% и 10%), т.к. значимость переменной «Численность населения с доходом ниже прожиточного минимума» находится на уровне 0,002 и ее коэффициент равняется 0,061, что больше нуля.

4) Четвертая гипотеза отвергается, т.к. уровень значимости переменной «Количество посещений музеев на 1000 человек» равняется 0,869. Кроме того, корреляция между «Количеством посещений музеев на 1000 человек» и «Количеством выпускников ВУЗов на 100 человек» отсутствует.

## 3.2. Оптимизированная линейная регрессия

При первой попытке построить линейную регрессию, мы столкнулись с тем, что некоторые переменные оказались незначимы. Начав по очереди исключать переменные из регрессии, нам удалось найти такой набор переменных, при которых все они были значимы на уровне 15%.



Стоит дать несколько пояснений относительно «дамми» переменных. Одна из таких переменных оказалась незначима в данной регрессии, однако значимы две другие. Однако невозможно исключить одну из «дамми» переменных, не исключив остальные, т.к. они все были получены из одной номинальной переменной. Соответственно, несмотря на «незначимость» «Большой плотности населения», мы решили оставить эту переменную в регрессии. Еще следует отметить, что из регрессии была автоматически исключена четвертая «дамми» переменная – «Плотность населения выше среднего», т.к. именно относительно нее были рассчитаны значения для других «дамми» переменных.

Кроме того, стоит отметить, что вместо обычного «Процента сбережений от получаемых доходов» был взят логарифм от данной переменной. Это объясняется тем, что между «Количеством преступлений» и «Процентом сбережений от получаемых доходов» наблюдается логарифмическая зависимость.

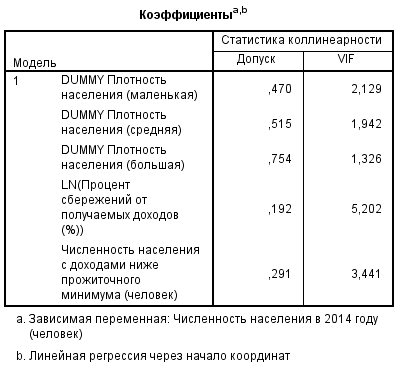
Таким образом, формула, определяющая количество совершенных преступлений может выглядеть следующим образом:

### 3.2.1. Тесты на гетероскедастичность и мультиколлинеарность

С переменными, которые входят в регрессию, мы провели ряд тестов на гетероскедастичность и мультиколлинеарность.

#### 3.2.1.1. VIF

Первый был VIF-тест на мультиколлинеарность. С помощью SPSS были построены таблицы следующего вида для всех переменных:



Приводить еще 5 подобных таблиц в нашем отчете мы не будем, лишь отметим, что во всех таблицах ни одного значения VIF, больше чем 10 не наблюдалось, что свидетельствует об отсутствии (или малой степени) мультиколлинеарности.

#### 3.2.1.2. Индекс обусловленности

В процессе выполнения этого теста была получена следующая матрица:



Далее, для определения индекса неопределённости необходимо найти собственные значения этой матрицы и разделить корень из максимального из них на корень из минимального из них.

В нашем случае это значение равно 36,416. Оно больше 30, поэтому по индексу обусловленности в данных присутствует мультиколлинеарность.

Исходя из двух тестов, которые дали противоположный результат, а также из того, что получившийся индекс обусловленности незначительно выше порогового значения, было решено не проводить над данными преобразований, призванных убрать мультиколлинеарность.

#### 3.2.1.3. Тест Парка

В данном тесте, проведенном на наших данных, значения p-value во всех регрессиях оказались больше, чем 0,05. Это говорит о том, что переменные не значимы, а, следовательно, гетероскедастичности нет.

#### 3.2.1.4. Тест Breusch-Pagan

В нашем случае при проверке значения на распределение ХИ-квадрат, мы получили 0,00055, что меньше необходимых 0,05. Это говорит о том, что значение не распределено как ХИ-квадрат, а значит – гетероскедастичность есть.

#### 3.2.1.5. Тест Goldfeld-Quandt

В нашем случае мы проделали данный тест три раза, используя различный независимые переменные. В двух случаях из трех полученной значение не было распределено в соответствии с F-распределением, что говорит о наличии гетероскедастичности.

### 3.2.2 Проверка гипотез для оптимизированной модели

1) Первая гипотеза значима на уровне значимости 1% (следовательно, и на уровне 5% и 10%), т.к. значимость переменной «Численность населения» находится на уровне 0,000 и ее коэффициент равняется 0,008, что больше нуля.

2) Вторая гипотеза отвергается на всех уровнях значимости меньше, чем 15%, т.к. значимость переменной «LN(Процент сбережений от получаемых доходов)» равняется 0,142.

3) Третья гипотеза, также как и первая, принимается на уровне значимости 1% (следовательно, и на уровне 5% и 10%), т.к. значимость переменной «Численность населения с доходом ниже прожиточного минимума» находится на уровне 0,003 и ее коэффициент равняется 0,056, что больше нуля.

4) Четвертая гипотеза отвергается на всех уровнях значимости, т.к. переменная «Количество посещений музеев на 100 человек» вообще не вошла в оптимизированную модель.

## 3.3. Сравнение начальной модели с оптимизированной

### 3.3.1. Модифицированный коэффициент детерминации

Для обеих моделей был выделен коэффициент детерминации он же для базовой модели он равен 0.83, в то время как для оптимизированной он уменьшился до значения в 0.827. Эти данные говорят о хорошей зависимости между рассматриваемыми переменными и правильности построения модели. Поскольку коэффициент нормирован к измененному числу переменных то факт незначительного уменьшения его величины не говорит об ухудшении качества модели.

### 3.3.2. Критерий Акаике (AIC)

Для базовой модели AIC был равен 1330, а для оптимизированной уменьшился до 1328, что говорит о возросшем качестве модели. Ввиду особенностей критерия, можно сделать вывод что значение уменьшилось ввиду уменьшения количества параметров, однако безусловно некую долю изменения внесло увеличение точности.

### 3.3.3. Критерий Шварца (BIC)

Также как и критерий Акаике, критерий Шварца уменьшился для оптимизированной модели по сравнению с базовой. Базовая модель – 1355, оптимизированная – 1344. Таким образом данный критерий также подтверждает выводы предыдущего: качество оптимизированной модели лучше, чем базовой.

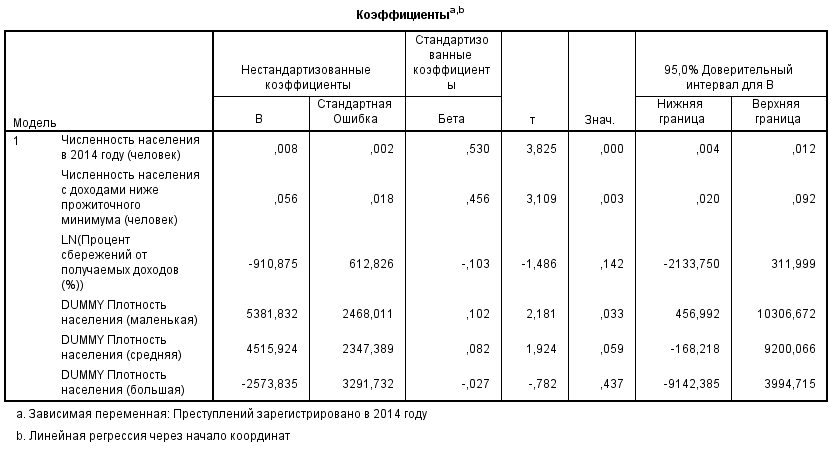
## 3.4 Вывод

После проведенных тестов можно сделать вывод, что оптимизированная модель позволяет лучше решить поставленный вопрос и обладает некоторыми преимуществами над базовой моделью.

# 4. Построение доверительных интервалов и проведение тестов на выбросы

## 4.1. Доверительные интервалы

Для построения доверительных интервалов был использован функционал программы SPSS. Вместе с коэффициентами и значимостью, для каждой переменной можно также вывести доверительные интервалы.



В данном случае доверительные интервалы необходимы для того, чтобы понять, в каких границах будет лежать коэффициент каждой переменной.

## 4.2. Тесты на выбросы

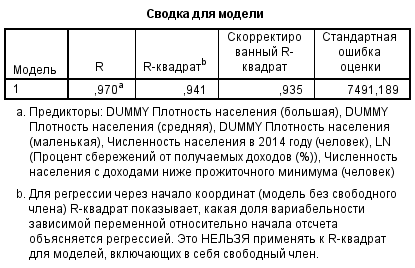
Для определения выбросов были использованы такие меры, как: «Расстояние Махаланобиса», «Расстояние Кука», «Ковариационное отношение», «DFFIT и стандартизованное DFFIT», «DFBETA и стандартизованное DFBETA». Исходя из имеющихся данных, были рассчитаны критические значения для каждой меры, а также значения каждой меры для каждого значения из выборки.

Результаты тестов (является значение выбросом или нет) были записаны в Excel-таблицу и отфильтрованы так, чтобы по крайней мере по трем тестам значения являлись выбросами. Таки значений было найдено всего 3, и они представлены на рисунке ниже:

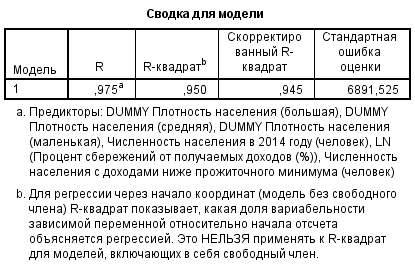




Данные значения было решено исключить из выборки и построить регрессию заново, сравнив полученный результат с предыдущим. В модели с выбросами был равен 0,941, в то время как в новой модели (без выбросов) это значение равно 0,950.



*Старая модель*



*Новая модель*

Очевидно, что новая модель лучше предыдущей и именно она будет использована, как финальная версия модели, разработанной в данном исследовании.

# 5. Заключение

В результате исследования было установлено, что между количеством совершаемых преступлений и различными социальными факторами действительно есть зависимость. Была составлена модель, предсказывающая количество преступлений на основе пяти других факторов. Большинство гипотез, выдвинутых в начале работы нашли свое подтверждение в процессе работы над моделью. Таким образом, данное исследование можно считать успешным.