**Первая часть кейса**

1)Удалить повторяющиеся строки из данных

2) Привести описательные статистики по переменным price и points. Минимально необходимые статистики:

* Число непустых значений
* Среднее значение
* Стандартное отклонение
* Медиана
* Мин. и макс. значения
* Размах

3) Отобразить гистограмму / плотность распределения для переменных points и price. Использовать пакет ggplot2 для визуализации. Минимум 2 графика.

4) Рассчитать корреляцию между переменными points и price.

* Отразить коэффициент корреляции и p-value
* Отразить доверительный интервал для коэффициента корреляции

5) Сформулировать комплексные выводы по результатам анализа.

Для этого ответьте на вопросы: по какой винодельне представлено наибольшее количество данных и что это означает; что характеризуют полученные описательные статистики для изучаемой страны, какая взаимосвязь прослеживается между переменными, и о чем это свидетельствует. Выводы должны быть не оборваны, а взаимосвязаны друг с другом - соблюдайте последовательность в их изложении. При необходимости постройте дополнительные графики.

**Решение**

Считываем данные:



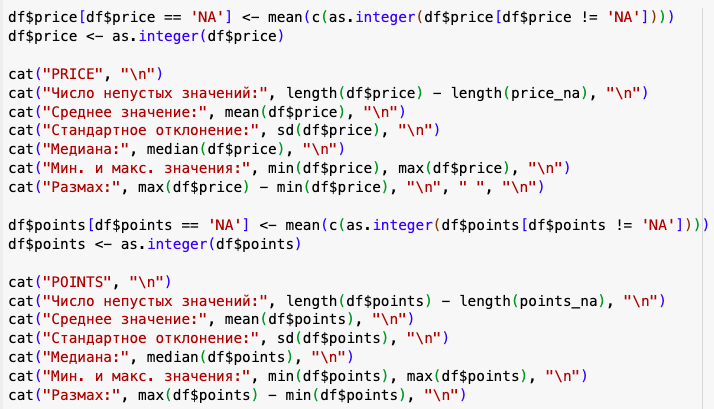
Удаляем повторы строк, оставляем только уникальные (1).



Запоминаем количество пустых строк для столбцов price и points



Приводим описательные статистики (2):

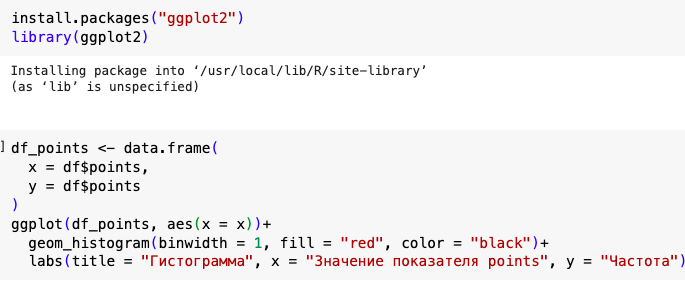


Вывод:

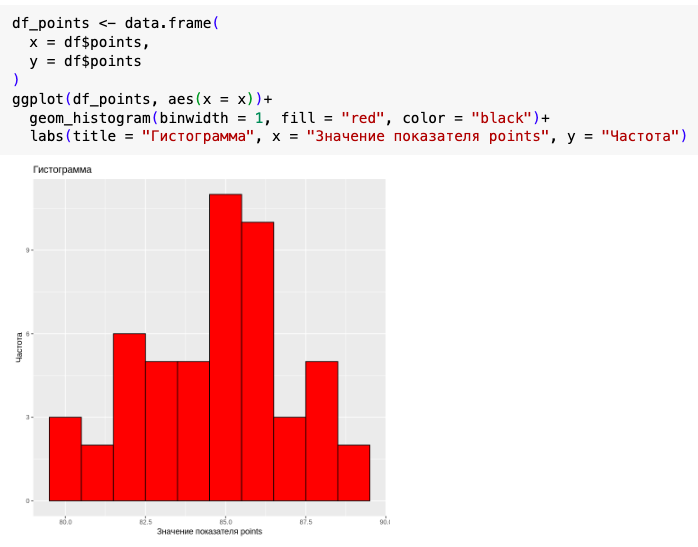
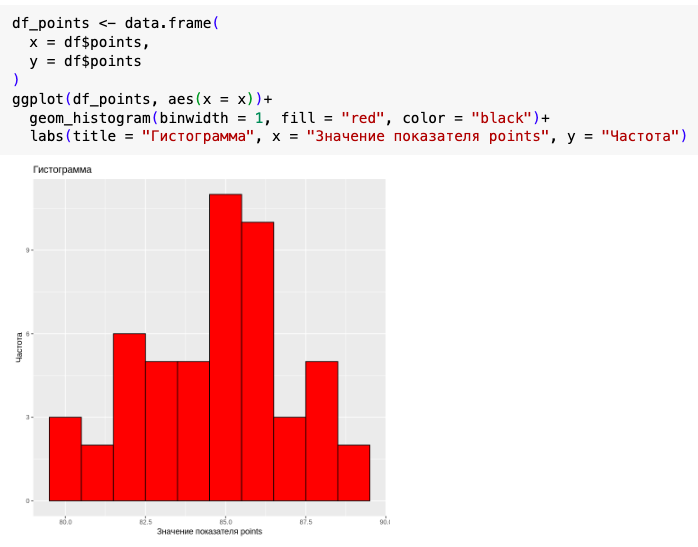


Данные по points демонстрируют меньшую изменчивость по сравнению с данными по price. Значения points сконцентрированы вокруг среднего значения, в то время как price характеризуется большим разбросом. Наличие большого размаха цен (price) может указывать на наличие выбросов.

Загружаем библиотеку ggplot2 для построения гистограмм (3).

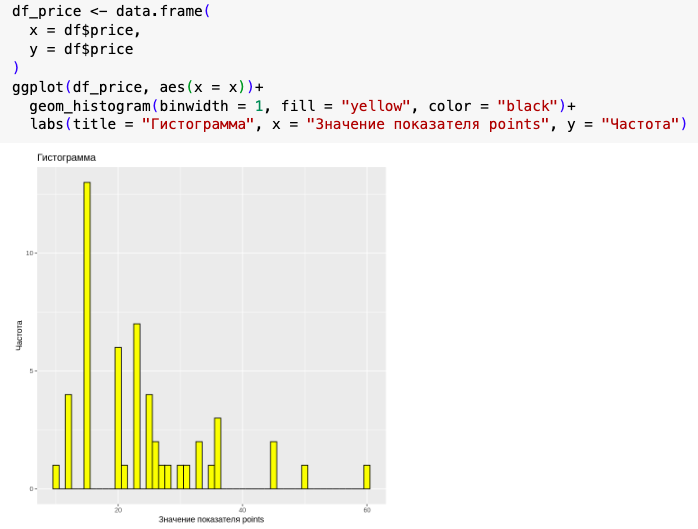


Гистограмма для показателя points



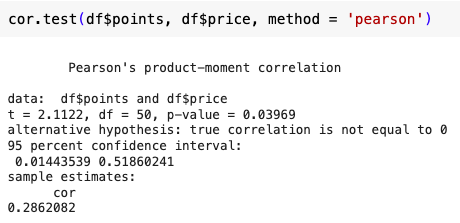
Гистограмма показывает распределение показателя points. Распределение можно охарактеризовать как приблизительно симметричное и близкое к нормальному, с модой в районе 86. Наблюдается небольшое количество значений на краях распределения, что может указывать на наличие незначительных выбросов или просто на естественную изменчивость данных.

Гистограмма для показателя price



Гистограмма показывает распределение значений переменной points. Видно, что распределение не является нормальным, а скорее имеет несколько пиков и значительный разброс значений. Наблюдается скопление значений в нескольких областях, а между ними есть промежутки с низкой частотой встречаемости.

Рассчитаем корреляцию между переменными points и price (4).

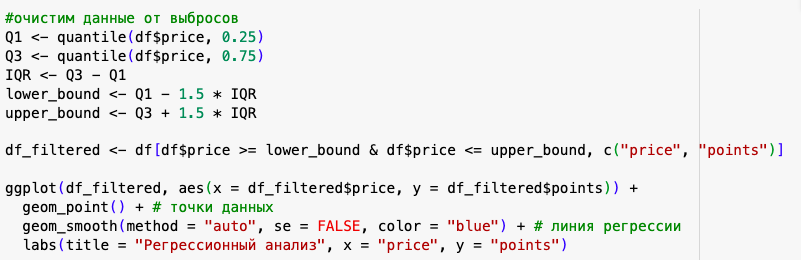


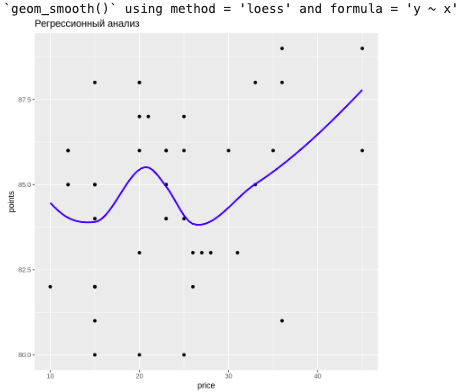
Коэффициент корреляции = 0.2862082

p-value = 0.03969

Доверительный интервал для коэффициента корреляции = 95%

Построим линию регрессии для дополнительных выводов.





**Комплексные выводы по результатам анализа (5)**

Больше всего данных об оценках критиками сортов вин представлено по винодельне «Casa Perini» (16 отзывов), это может быть связано с тем, что у них представлено самое большое количество сортов вина (11 видов).

Средняя цена за бутылку вина из Бразилии составляет 23,7 д.ед., размах цен составляет 50 д.ед. Потребители могут приобрести как дорогие сорта вин из Бразилии, максимальная цена которых составляет 60 д.ед., так и более бюджетные варианты с ценой 10 д.ед.

Среднее значение оценки критиком Бразильского вина равно 84,7 пункта. Поскольку в базе данных хранятся только оценки не ниже 80 баллов, среднее значение, рассчитанное по этим данным, завышено по сравнению со средним значением всех оценок (включая оценки ниже 80 баллов). Настоящее среднее значение всех оценок бразильского вина будет ниже 84,7 баллов. Насколько ниже — сказать невозможно без информации об оценках ниже 80 баллов и их количестве.

Значение p-value (0.0397) меньше уровня значимости 0.05, что позволяет отвергнуть нулевую гипотезу об отсутствии корреляции между этими переменными. Таким образом, можно заключить, что существует связь между оценкой points и ценой price, при повышении одной из них, другая также имеет тенденцию к повышению. Однако, сила этой связи невысока, о чем свидетельствует относительно небольшое значение коэффициента корреляции (0.286). 95% доверительный интервал (0.014; 0.519) указывает на неопределенность оценки силы связи.

График регрессии демонстрирует нелинейную зависимость между ценой (price) и оценкой (points). Наблюдается немонотонная связь: с ростом цены оценка сначала растёт, достигает пика, а затем снова падает, после чего снова начинает расти. Это говорит о том, что высокая цена не гарантирует высокую оценку.

**Вторая часть кейса**

|  |  |  |
| --- | --- | --- |
| Фамилия студента | Файл wine\_reviews | Модель регрессии |
| Безгребельная | Бразилия | У от Х2Х4 |

1. Оцените выбросы в распределении переменных, удалите их, если они есть и составляют малую долю в данных (вы можете удалить выбросы выборочно). Проверьте однородность переменных с помощью правила трех сигм/межквартильного размаха до и после удаления выбросов. Сделайте выводы.
2. Оцените параметры полученной модели по всем ее характеристикам. Ответьте на вопросы: значимо ли уравнение регрессии? Значимы ли отдельные коэффициенты уравнения регрессии? Что это означает? Постройте 2 графика зависимости У от зависимых переменных попарно.
3. **Сформулируйте комплексные выводы по результатам моделирования**. Можно ли использовать полученную модель? Почему да/нет? Приведите экономическую интерпретацию модели, даже если она некорректная.

Загружаем данные.



# Создаем подмножество столбца x2, удаляя выбросы

df\_x2 <- df$x2[as.numeric(df$x2) <= quantile(as.numeric(df$x2), 0.75) + IQR(df$x2)\*1.5 & as.numeric(df$x2) >= quantile(as.numeric(df$x2), 0.25) - 1.5\*IQR(df$x2)]

# Создаем подмножество столбца x4, удаляя выбросы

df\_x4 <- df$x4[as.numeric(df$x4) <= quantile(as.numeric(df$x4), 0.75) + IQR(df$x4)\*1.5 & as.numeric(df$x4) >= quantile(as.numeric(df$x4), 0.25) - 1.5\*IQR(df$x4)]

# Создаем ящичную диаграмму для исходного x2

plot2 <- ggplot(df, aes(y = x2)) +

geom\_boxplot(fill = "red", color = "black") +

labs(title = "Ящичная диаграмма x2", y = "x2")+

theme\_minimal()

# Создаем ящичную диаграмму для исходного x4

plot4 <- ggplot(df, aes(y = x4)) +

geom\_boxplot(fill = "blue", color = "black") +

labs(title = "Ящичная диаграмма x4", y = "x4")+

theme\_minimal()

# Создаем датафреймы для x2 и x4 после удаления выбросов (df\_x2 и df\_x4 уже содержат очищенные данные)

df2 <- data.frame(

x = df\_x2,

y = df\_x2

)

df4 <- data.frame(

x = df\_x4,

y = df\_x4

)

# Создаем ящичную диаграмму для x2 после удаления выбросов

plot2\_new <- ggplot(df2, aes(y = x)) +

geom\_boxplot(fill = "red", color = "black") +

labs(title = "Ящичная диаграмма x2", y = "x2") +

theme\_minimal()

# Создаем ящичную диаграмму для x4 после удаления выбросов

plot4\_new <- ggplot(df4, aes(y = x)) +

geom\_boxplot(fill = "blue", color = "black") +

labs(title = "Ящичная диаграмма x4", y = "x4") +

theme\_minimal()

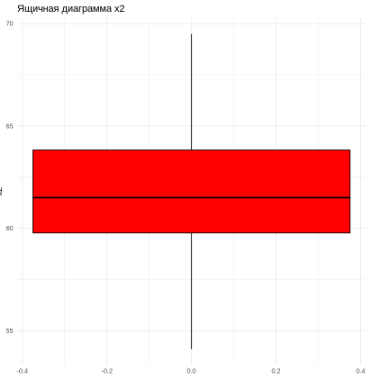
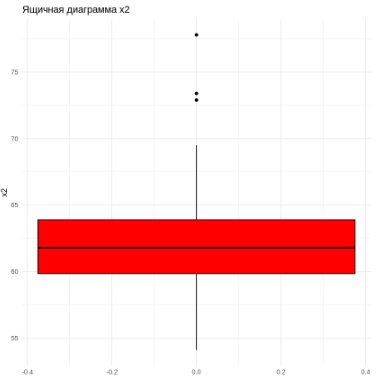
# Отображаем все четыре графика

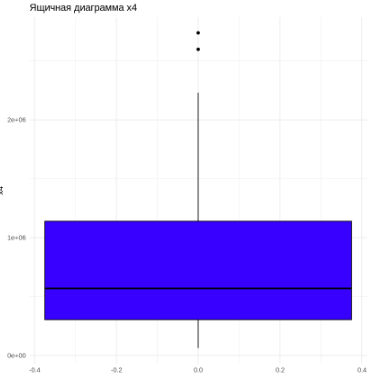
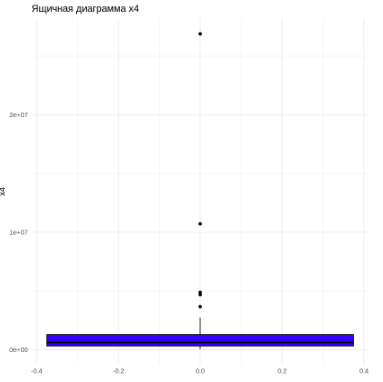
plot2

plot2\_new

plot4

plot4\_new





Проверим однородность переменных с помощью межквартильного размаха

#IQR до удаления выбросов

q1\_before\_2 = quantile(df$x2, 0.25, na.rm = TRUE) # Первый квартиль x2 до очистки

q3\_before\_2 = quantile(df$x2, 0.75, na.rm = TRUE) # Третий квартиль x2 до очистки

iqr\_before\_2 = q3\_before\_2 - q1\_before\_2 # IQR x2 до очистки

q1\_before\_4 = quantile(df$x4, 0.25, na.rm = TRUE) # Первый квартиль x4 до очистки

q3\_before\_4 = quantile(df$x4, 0.75, na.rm = TRUE) # Третий квартиль x4 до очистки

iqr\_before\_4 = q3\_before\_4 - q1\_before\_4 # IQR x4 до очистки

#IQR после удаления выбросов

q1\_after\_2 <- quantile(df\_x2, 0.25, na.rm = TRUE)

q3\_after\_2 <- quantile(df\_x2, 0.75, na.rm = TRUE)

iqr\_after\_2 <- q3\_after\_2 - q1\_after\_2

q1\_after\_4 <- quantile(df\_x4, 0.25, na.rm = TRUE)

q3\_after\_4 <- quantile(df\_x4, 0.75, na.rm = TRUE)

iqr\_after\_4 <- q3\_after\_4 - q1\_after\_4

iqr\_before\_2

iqr\_after\_2

iqr\_before\_4

iqr\_after\_4

Вывод:

**75%:** 4.05

**75%:** 4.05

**75%:** 969765.235

**75%:** 835356.424

Значение межквартильного размаха для х4 изменилось с 969 765.235 до

835 356.424 (на 14%), это значительное снижение, указывает на то, что выбросы существенно влияли на разброс данных в центральной части распределения. Удаление этих выбросов привело к более компактному и однородному распределению данных. Это подтверждается изменениями среднего и медианы.

cat("Среднее х4 до удаления выбросов: ", mean(df$x4), "\n")

cat("Среднее х4 после удаления выбросов: ", mean(df\_x4), "\n")

cat( "\n")

cat("Медиана х4 до удаления выбросов: ", median(df$x4), "\n")

cat("Медиана х4 после удаления выбросов: ", median(df\_x4), "\n")

Среднее х4 до удаления выбросов: 1359037

Среднее х4 после удаления выбросов: 763596.8

Медиана х4 до удаления выбросов: 619689.7

Медиана х4 после удаления выбросов: 569747.5

Удаление выбросов из переменной x4 существенно повлияло на среднее значение, снизив его с 1359037 до 763596.8, это указывает на то, что выбросы имели сильное влияние на среднее значение. Медиана, в отличие от среднего, изменилась менее значительно. В целом, результаты свидетельствуют о наличии значительных выбросов в исходных данных x4, которые были успешно удалены.

2) Оцените параметры полученной модели по всем ее характеристикам. Ответьте на вопросы: значимо ли уравнение регрессии? Значимы ли отдельные коэффициенты уравнения регрессии? Что это означает? Постройте 2 графика зависимости У от зависимых переменных попарно.

Построим модель регрессии для очищенных от выбросов значений.

# Возвращает TRUE или FALSЕ в зависимости от того, входит ли значение в межквартильных размах или нет

remove\_outliers\_iqr <- function(x) {

q1 <- quantile(x, 0.25, na.rm = TRUE)

q3 <- quantile(x, 0.75, na.rm = TRUE)

iqr <- q3 - q1

lower\_bound <- q1 - 1.5 \* iqr

upper\_bound <- q3 + 1.5 \* iqr

x >= lower\_bound & x <= upper\_bound

}

outlier\_indices\_x2 <- !remove\_outliers\_iqr(df$x2)

outlier\_indices\_x4 <- !remove\_outliers\_iqr(df$x4)

# Комбинирует индексы выбросов (если есть строка с выбросом (либо Х2, либо Х4), то она удаляется)

combined\_outlier\_indices <- outlier\_indices\_x2 | outlier\_indices\_x4

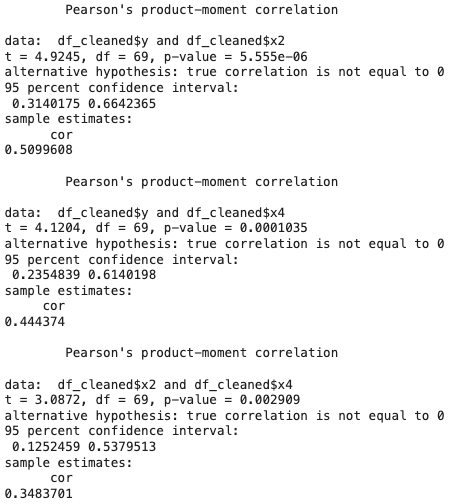
# Удаляются строки с выбросами из трех столбцов

df\_cleaned <- df[!combined\_outlier\_indices, c("y", "x2", "x4")]

cor.test(df\_cleaned$y, df\_cleaned$x2, method = 'pearson')

cor.test(df\_cleaned$y, df\_cleaned$x4, method = 'pearson')

cor.test(df\_cleaned$x2, df\_cleaned$x4, method = 'pearson')



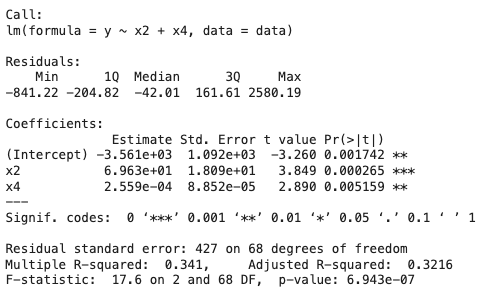
|  |  |  |
| --- | --- | --- |
| Корреляция между x2 и y  (r = 0.51) | Корреляция между x4 и y  (r = 0.44) | Корреляция между x2 и x4  (r = 0.35) |
| Указывает на умеренную положительную связь между этими переменными. При увеличении x2 тенденция к увеличению y. | Также демонстрирует умеренную положительную связь. При увеличении x4 наблюдается тенденция к увеличению y. | Показывает слабую, но всё же заметную положительную корреляцию между x2 и x4. |

Все три корреляции (y с x2, y с x4, и x2 с x4) статистически значимы (p-значения намного меньше 0.05). Это означает, что существует статистически значимая связь между парами переменных.

model <- lm(y ~ x2 + x4, data = data)

#выводим параметры

summary(model)



Результаты регрессионного анализа показывают, что оба коэффициента, x2 и x4, статистически значимы. Это видно по p-значениям (Pr(>|t|)) для этих коэффициентов: 0.000265 и 0.005159 соответственно. Оба значения значительно меньше общепринятого уровня значимости 0.05.

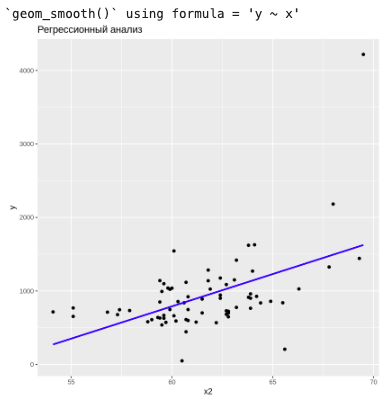
#строим график y от x2

ggplot(data, aes(x = x2, y = y)) +

geom\_point() + # точки данных

geom\_smooth(method = "lm", se = FALSE, color = "blue") + # линия регрессии

labs(title = "Регрессионный анализ", x = "x2", y = "y")



Линия регрессии имеет положительный наклон, что подтверждает наличие положительной корреляции.

#строим график y от x4

ggplot(data, aes(x = x4, y = y)) +

geom\_point() + # точки данных

geom\_smooth(method = "lm", se = FALSE, color = "red") + # линия регрессии

labs(title = "Регрессионный анализ", x = "x4", y = "y")

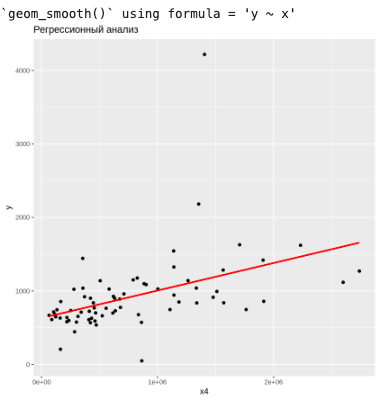


График рассеяния с линией регрессии демонстрирует слабую положительную связь между переменными x4 и y. Линия регрессии имеет положительный наклон, что указывает на тенденцию к увеличению y с ростом x4. Однако, рассеяние точек вокруг линии регрессии достаточно велико, что свидетельствует о невысокой силе этой связи.

**Комплексные выводы**

Оба независимых показателя, *x2* и *x4*, имеют статистически значимые коэффициенты (p-значения < 0.05). Это означает, что влияние как *x2*, так и *x4* на *y* не случайно.

Высокая F-статистика (17.6) с низким p-значением (6.943e-07) подтверждает, что как минимум один из предикторов значимо влияет на y. Модель в целом значима.

R-квадрат (0.341) указывает, что модель объясняет 34.1% вариации в *y*. Это относительно низкое значение, что говорит о том, что существуют другие факторы, влияющие на *y,* которые не включены в модель.

Использовать модель можно, но прогнозы, основанные на этой модели, будут иметь значительную погрешность, т.к. низкий R-квадрат свидетельствует о том, что модель объясняет лишь небольшую часть вариации в *y.*

**Экономическая интерпретация:**

Y – представляет какой-то экономический показатель на уровне региона, например, ВРП на душу населения.

Х2 – объем промышленного производства.

Х4 – показатель численности населения.

Модель описывает зависимость ВРП на душу населения (Y) от объема промышленного производства (X2) и численности населения (X4).