МИНОБРНАУКИ РОССИИ

РГУ НЕФТИ И ГАЗА (НИУ) ИМЕНИ И.М. ГУБКИНА

ФАКУЛЬТЕТ АВТОМАТИКИ И ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ

КАФЕДРА ИНФОРМАТИКИ

ДИСЦИПЛИНА «ОСНОВЫ АНАЛИЗА БОЛЬШИХ ДАННЫХ И МАШИННОЕ ОБУЧЕНИЕ»

**О Т Ч Е Т**

**по Домашнему Заданию 2**

**«Анализ данных в среде R»**

Выполнила: студентка группы АА-19-05

Данилова М.А.

Проверила: доцент Вишневская Е. А.

Москва 2022

**Выбрать тему из предлагаемого списка: 9. Регрессионный анализ**

**Регрессионный анализ: основные положения.**

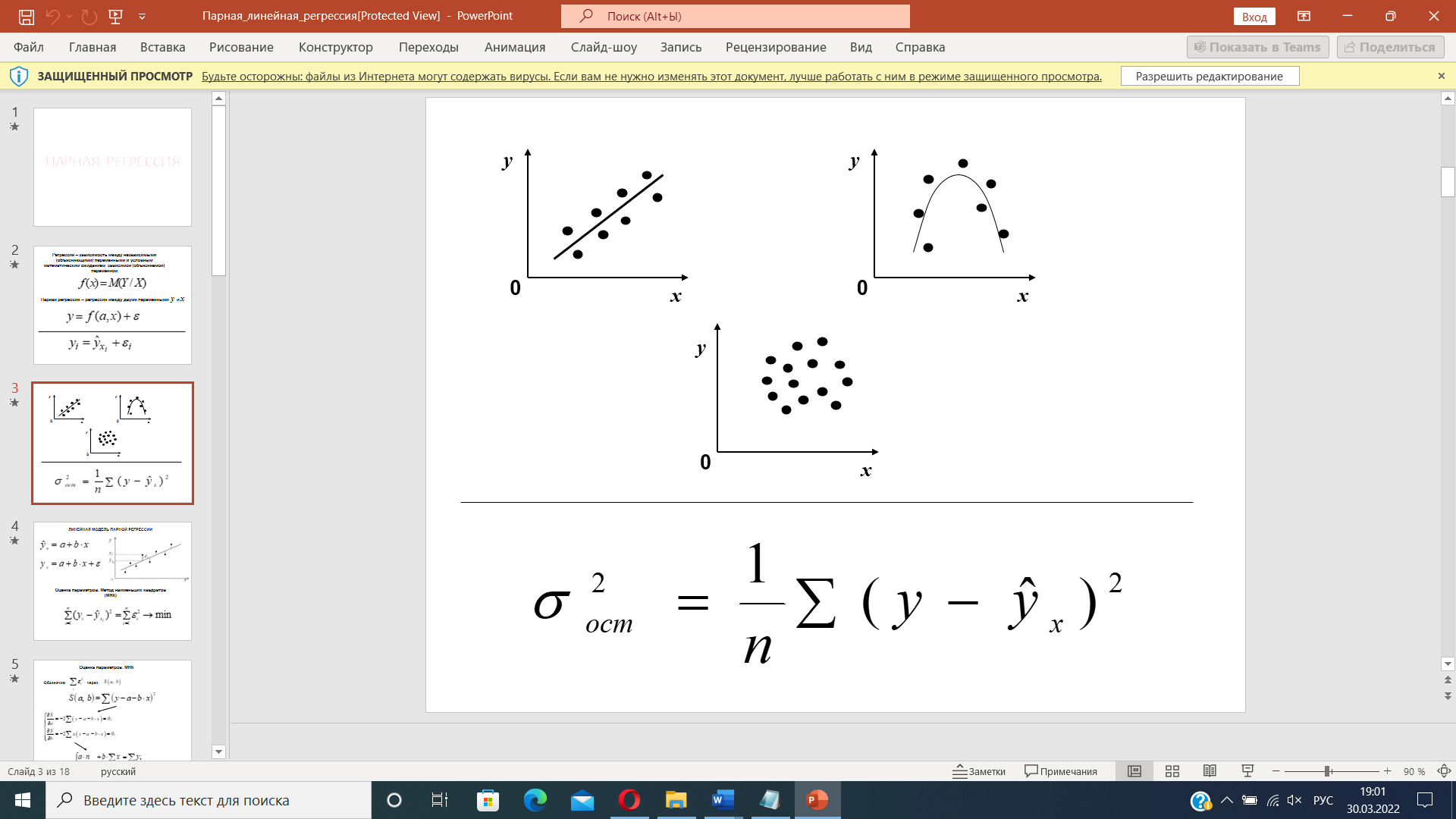
Регрессионные модели используются для прогнозирования непрерывных целевых значений (прогнозирования цен на жилье, прогноза погоды). Под регрессией понимается функциональная зависимость между объясняющими переменными xi и условным математическим ожиданием (средним значением) зависимой переменой y; модель строится с целью прогнозирования этого среднего значения при фиксированных значениях первых. Любая регрессионная модель позволяет обнаружить только количественные зависимости, которые не обязательно отражают причинные. В ходе регрессионного анализа определяют коэффициенты регрессии (β) - величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

В общем случае, предположим, что мы наблюдаем количественный отклик Y, и несколько разных предикторов X1, X2 , . . .,Xp . Предположим, что есть какая-то взаимосвязь между Y и X = (X1, X2 ,...,Xp ), которая в общей форме может быть записана в виде

Y = f(X) + є.

Здесь f – это некоторая фиксированная неизвестная функция переменных X1, X2 ,...,Xp , и є – случайная ошибка, не зависящая от X, с нулевым матожиданием.

Выбор формулы связи переменных называется **спецификацией** уравнения регрессии. В случае парной регрессии выбор формулы обычно осуществляется по графическому изображению реальных статистических данных.

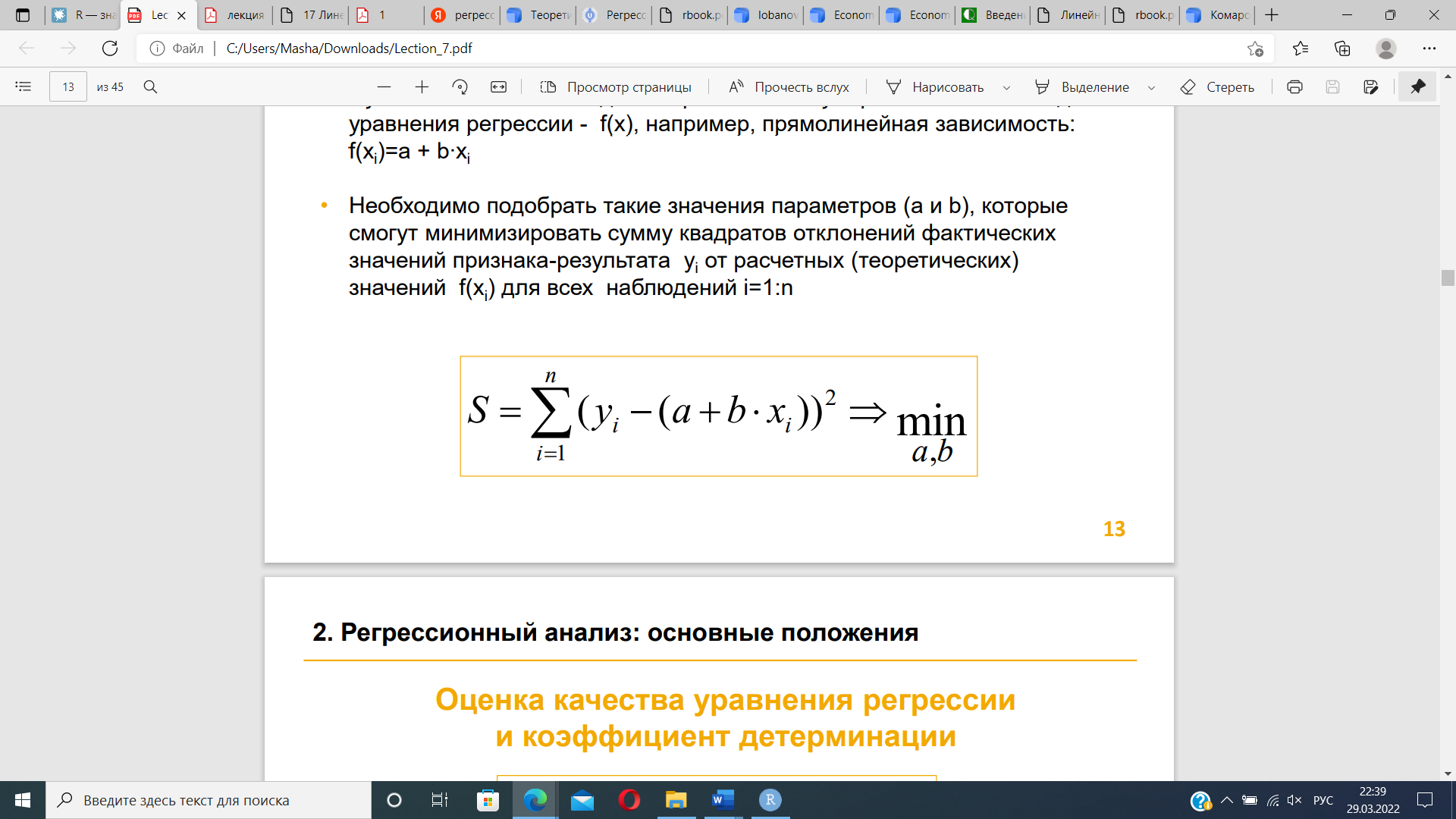
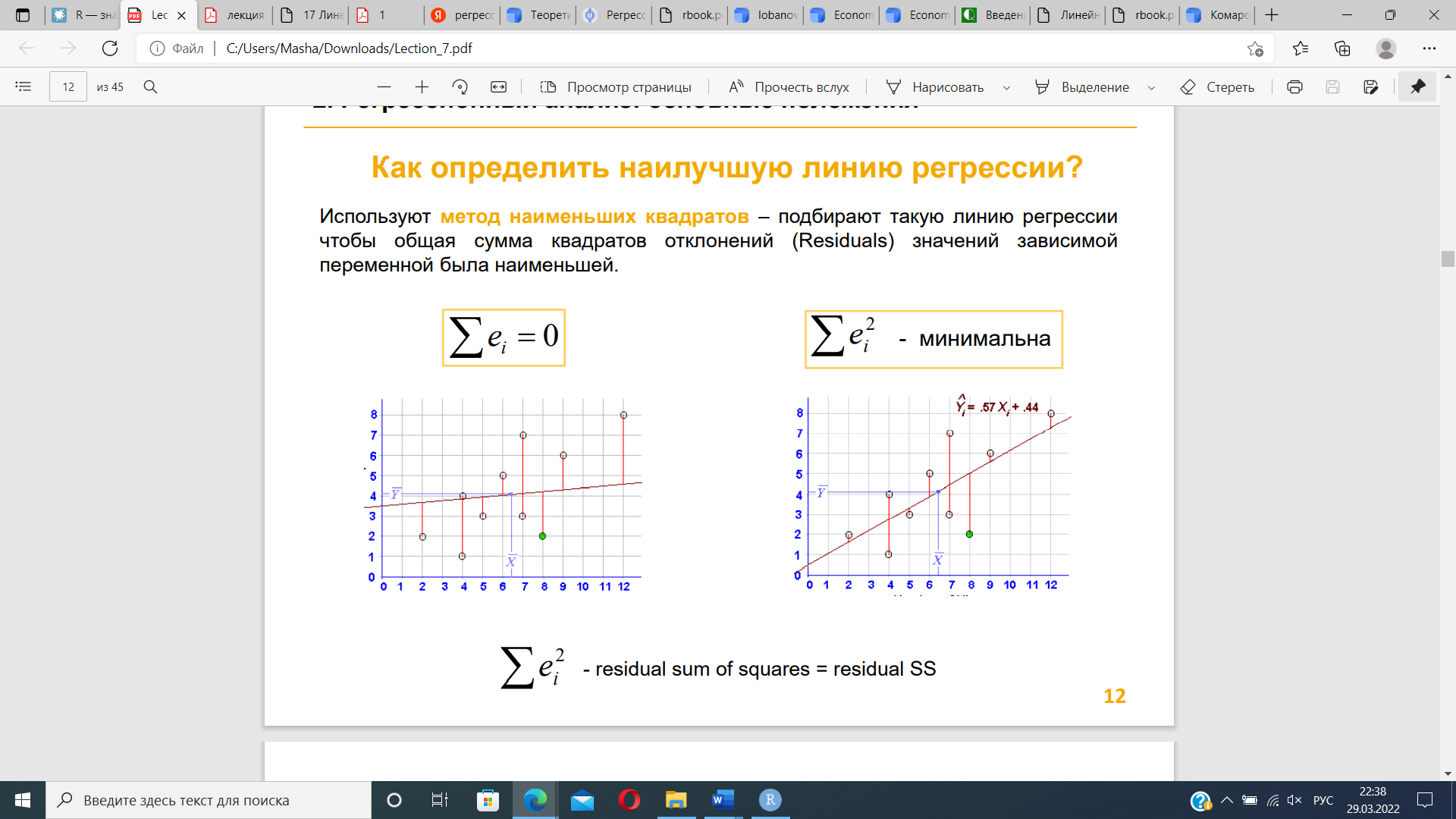


**Парная (простая) линейная регрессия.** Этот подход для прогнозирования количественного отклика Y на основе единственной предикторной переменной X. Предполагается, что есть приблизительная линейная взаимосвязь между Y и X. Математически можно записать эту взаимосвязь следующим образом:

Yi = a + bXi, где

Yi – зависимая переменная и Xi – независимая переменная, a – константа, b – угловой коэффициент, характеризует наклон прямой и показывает, на какую величину в среднем изменится результативный признак Yi, если переменная Xi увеличится на единицу своего измерения.

Для определения наилучшей линии регрессии используют **метод наименьших квадратов**, те добиваются, чтобы сумма квадратов остатков e была минимальной. Под остатками понимается разность между очередным наблюдением и прогнозом модели.



В общем случае для n наблюдений решают систему уравнений:



Толковой интерпретации регрессионных коэффициентов мешает также различие в единицах измерения. Например, если предиктор измеряется в сантиметрах, его вес

будет в 100 раз отличаться по весу от предиктора, берущегося в метрах. Чтобы избежать такого, мы должны **стандартизировать** единицы измерения предикторных переменных перед тем, как проводить регрессионный анализ. Стандартизация – это выражение переменных в процентилях.

При использовании линейной регрессии в качестве показателем тесноты связи выступает линейный **коэффициент корреляции** (чем ближе коэффициент по модулю к единице, тем теснее связь).

**Множественная регрессия** является расширением простой линейной регрессии. Она исследует влияние двух и более предикторов на критерий (Y=B1\*X1+B2\*X2+B3\*X3+…+A).

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели, который включает 2 круга вопросов: отбор факторов и выбор уравнения регрессии.

**Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:**

1. Они должны быть **количественно измеримы**. Если необходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность.

2. Каждый **фактор должен быть достаточно тесно связан с результатом** (т.е. коэффициент парной линейной корреляции между фактором и результатом должен быть существенным).

3. **Факторы не должны быть сильно коррелированы друг с другом** или находиться в строгой функциональной связи (т.е. они не должны быть интеркоррелированы). Мультиколлинеарность может привести к нежелательным последствиям. Существуют различные подходы преодоления сильной межфакторной корреляции. Простейший из них – исключение из модели факторов, в наибольшей степени ответственных за мультиколлинеарность. Определение факторов, ответственных за мультиколлинеарность, может быть основано на анализе матрицы межфакторной корреляции. При этом определяют пару признаков-факторов, которые сильнее всего связаны между собой (коэффициент линейной парной корреляции максимален по модулю). Из этой пары в наибольшей степени ответственным за мультиколлинеарность будет тот признак, который теснее связан с другими факторами модели (имеет более высокие по модулю значения коэффициентов парной линейной корреляции).

**Коэффициенты VIF** (variance inflation factor) показывают, насколько сильно связаны друг с другом регрессоры модели. Если коэффициенты VIF для всех регрессоров оказались меньше 10 (иногда используют 5), это значит, что существенной мультиколлинеарности в модели не наблюдается. В противном случае стоит сделать вывод о том, что в модели есть мультиколлинеарность.

4. **Отсутствие автокорреляции** – отсутствие независимости остатков. Выявляется с помощью теста Дурбина-Уотсона (обнаруживает автокорреляцию первого порядка).

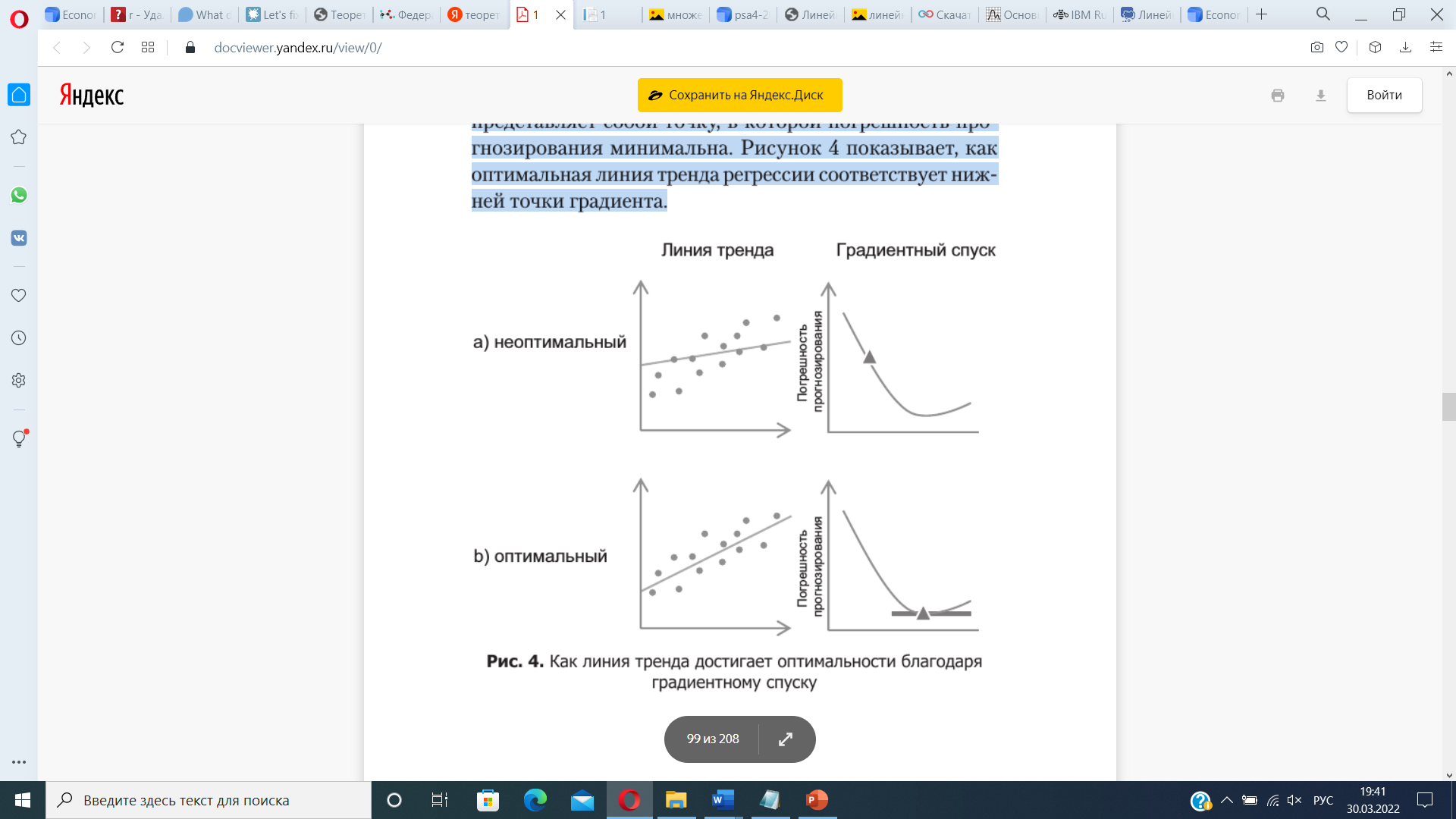
‒ Если d=2 – отсутствие автокорреляции.

При выборе формы уравнения множественной регрессии предпочтение отдается линейной функции в виду четкой интерпретации параметров. Параметры уравнения множественной регрессии можно также оценить методом наименьших квадратов, составив и решив систему нормальных линейных уравнений.

**Градиентный спуск (gradient descent)** используется в случаях, когда параметры уравнения нельзя получить путем решения систем уравнений. Алгоритм градиентного спуска делает первоначальное предположение о наборе весовых составляющих, после чего начинается итеративный процесс их применения к каждому элементу данных для прогнозирования, а затем они перенастраиваются для снижения общей ошибки прогнозирования.

Этот процесс можно сравнивать с пошаговым спуском в овраг в поисках дна. На каждом этапе алгоритм определяет, какое направление даст наиболее крутой спуск,

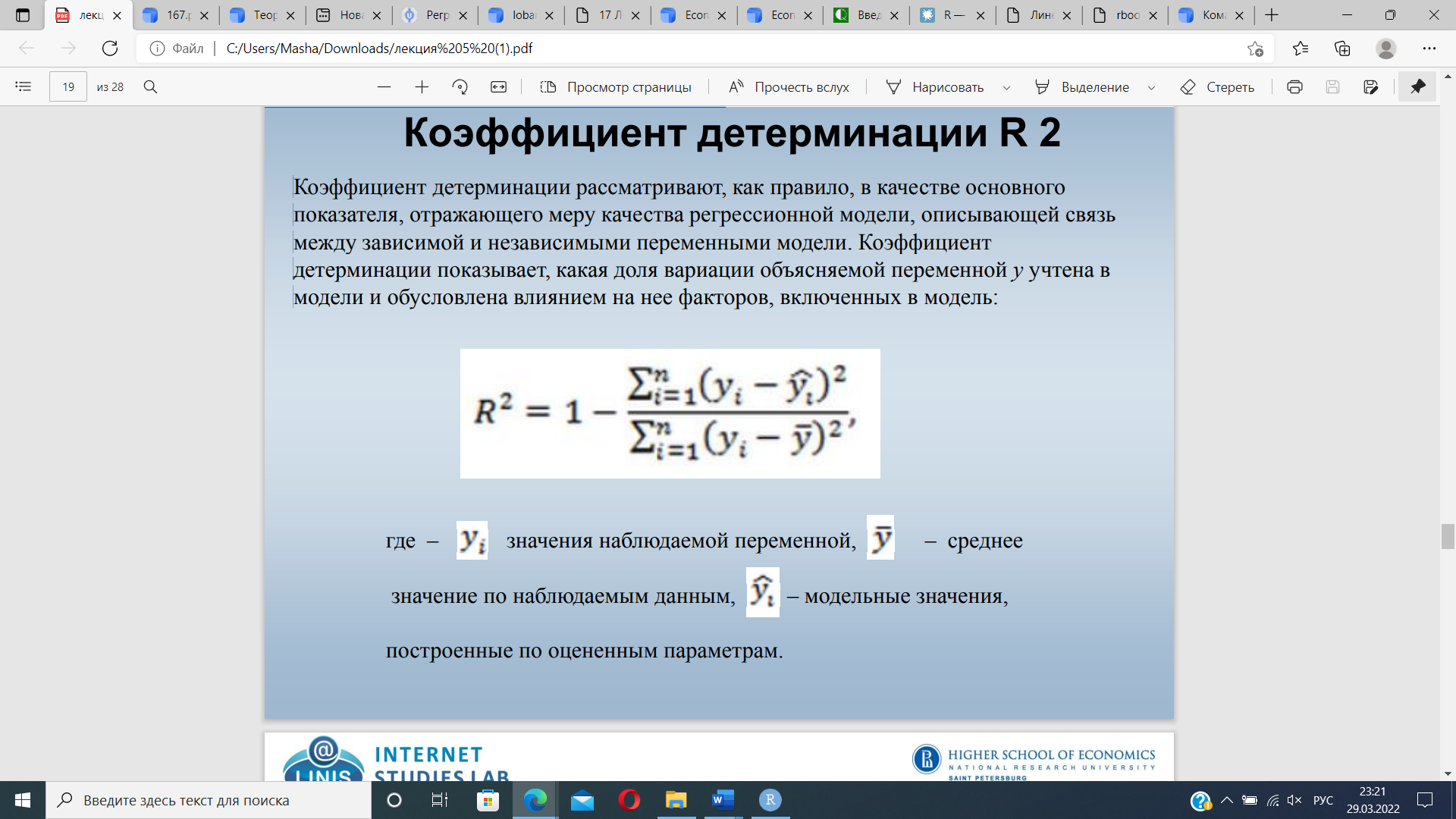
и пересчитывает весовые составляющие. В конечном итоге мы достигнем самой нижней позиции, которая представляет собой точку, в которой погрешность прогнозирования минимальна. Рисунок показывает, как оптимальная линия тренда регрессии соответствует нижней точки градиента.



Кроме регрессии градиентный спуск может также использоваться для оптимизации параметров в других моделях, таких как метод опорных векторов или в нейронных сетях.

**Оценка качества уравнения регрессии.**

Коэффициент детерминации рассматривают в качестве основного показателя, отражающего меру качества регрессионной модели. Он показывает, какая доля вариации объясняемой переменной y учтена в модели и обусловлена влиянием на нее факторов, включенных в модель:



- значения наблюдаемой переменной, – среднее значение по наблюдаемым данным, – модельные значения, построенные по оцененным параметрам.

Чем ближе R-квадрат к 1, тем выше качество регрессионной модели (факторы сильнее влияют на результат).

**Значимость уравнения регрессии и отдельных параметров.**

Проверить значимость уравнения регрессии – значит установить, соответствует ли

аналитическая модель экспериментальным данным, и достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной. Оценка значимости уравнения регрессии в целом производится на основе **F-критерия Фишера** (чем больше значение параметра — тем лучше).

Для проверки значимости коэффициента регрессии **применяется t -распределение Стьюдента** (если есть основания считать, что между величинами Y и X нет линейной зависимости, то коэффициент статистически незначим-слишком близок к 0).

Если между изучаемыми явлениями существуют нелинейные соотношения, то

они выражаются с помощью соответствующих нелинейных функций (полиномы различных степеней, гипербола, степенная, показательная, экспоненциальная регрессии и тд).

**Для сравнения регрессионных моделей по степени точности предсказаний используются метрики оценки.**

***MSE (Mean Squared Error)*** *и*змеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех точек данных. Самая популярная метрика, используемая для задач регрессии. Усиливается влияние ошибок по квадратуре от исходного значения.

[](http://4.bp.blogspot.com/-8-nlPIjzBEU/Wuq14EZbqAI/AAAAAAAACZY/XLScxPJAbBIpY8fEc_YZuW7wR0QxYqfhQCK4BGAYYCw/s1600/MSE.png)

Чем меньше MSE, тем точнее наше предсказание. Оптимум достигается в точке 0. Является дифференцируемой, что позволяет более эффективно использовать для поиска экстремумов с помощью математических методов.

***Root Mean Squared Error (RMSE)*** *- к*орень от квадратной ошибки. Ее легко интерпретировать, поскольку она имеет те же единицы, что и исходные значения (в отличие от MSE). Также она оперирует меньшими величинами по абсолютному значению, что может быть полезно для вычисления на компьютере.

[](http://4.bp.blogspot.com/-SuftZtem2_E/Wuq152CYCXI/AAAAAAAACZg/o1hLEvFE_Cgf3JEdVZiMvlloK_LFKvrwQCK4BGAYYCw/s1600/RMSE.png)

Итак, **основными этапами регрессионного анализа являются:**

1. Выбор вида уравнения регрессии (спецификация модели).

2. Выбор независимых переменных, оказывающих существенное влияние на

зависимую переменную.

3. Оценка параметров уравнения регрессии (параметризация модели).

4. Оценка статистической надежности регрессионной модели (верификация).

**Данные с подходящей структурой для выбранного метода**.

Для анализа используем данные автомобильной компании Geely Auto, представленные на сайте Kaggle: <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction>

Задача будет состоять в том, чтобы определить взаимосвязь между различными параметрами автомобилей и их ценой на рынке.

Столбцы таблицы (все данные-целые числа):

car\_ID – ид автомобиля

fueltype – тип топлива (1=газ, 2=дизельное)

aspiration – ускорение (стандарт, турбо)

drivewheel – ведущее колесо (переднее или заднее)

wheelbase – база шасси

carlength – длина авто

carwidth – ширина авто

carheight – высота авто

curbweight – снаряженная масса

enginetype – тип двигателя

cylindernumber – количество цилиндров

enginesize – размер двигателя

boreratio – коэффициент проходимости

horsepower – число лошадиных сил

peakrpm – пиковые обороты

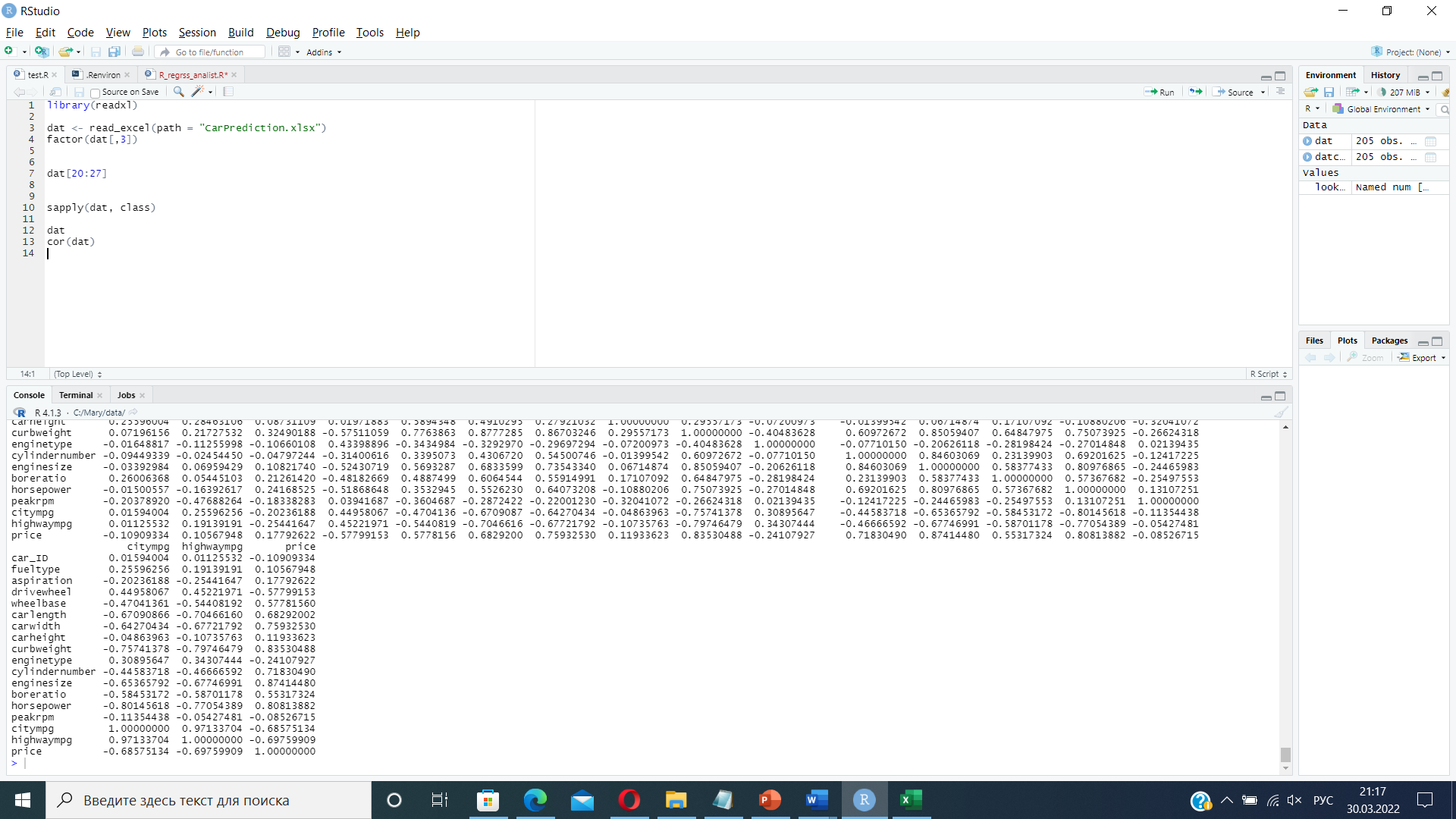
citympg – ситимиль на галлон

highwaympg – шоссемиль на галлон

price – цена, целевая ячейка

Для анализа в среде R потребуется RStudio, а также инструмент RTools.

Рабочее поле RStudio выглядит следующим образом:



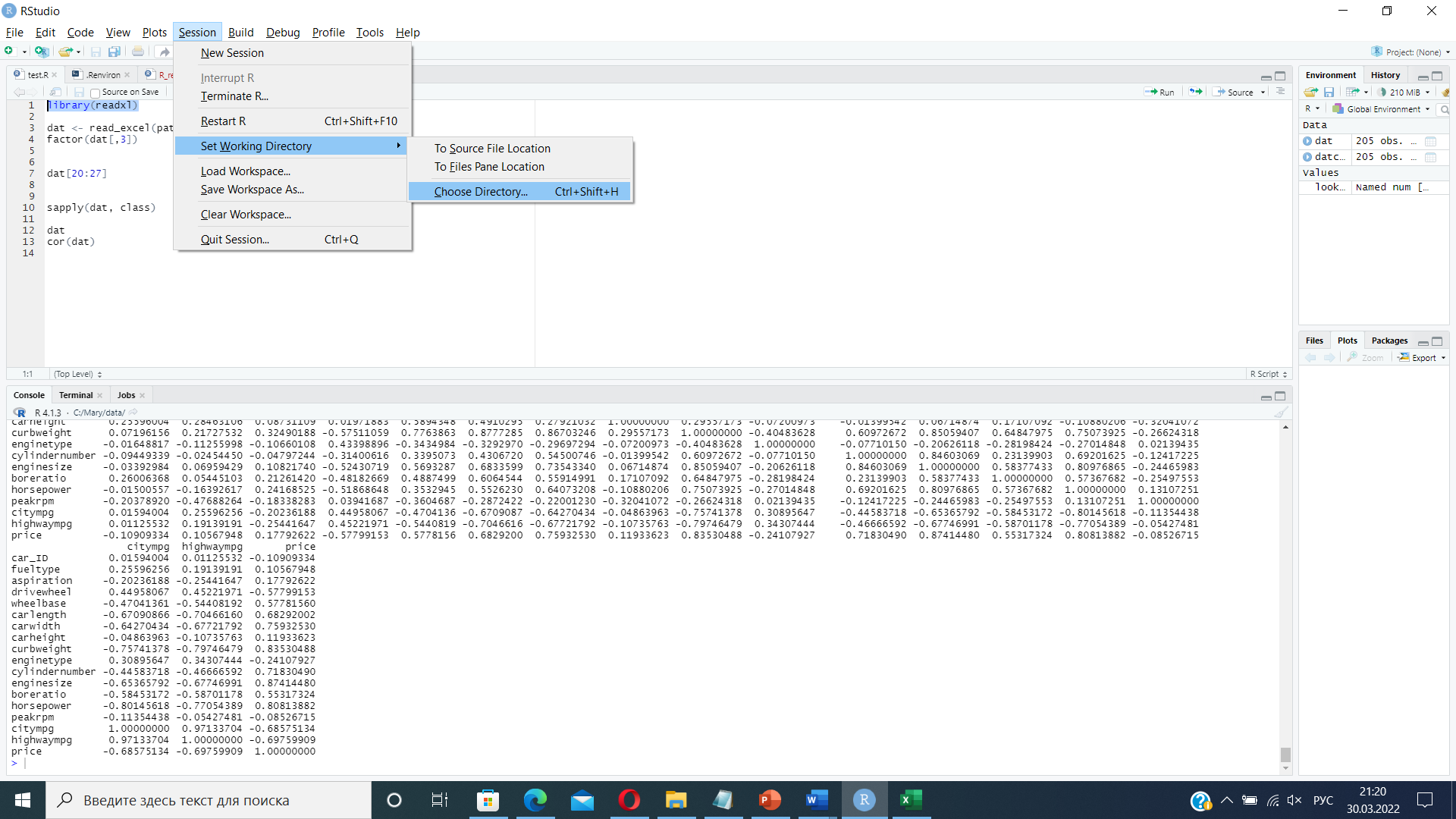
Этап 1. Чтение данных.

Для чтения данных из Excel установим необходимый пакет

[install.packages](https://rdrr.io/r/utils/install.packages.html)("readxl")

library(readxl)

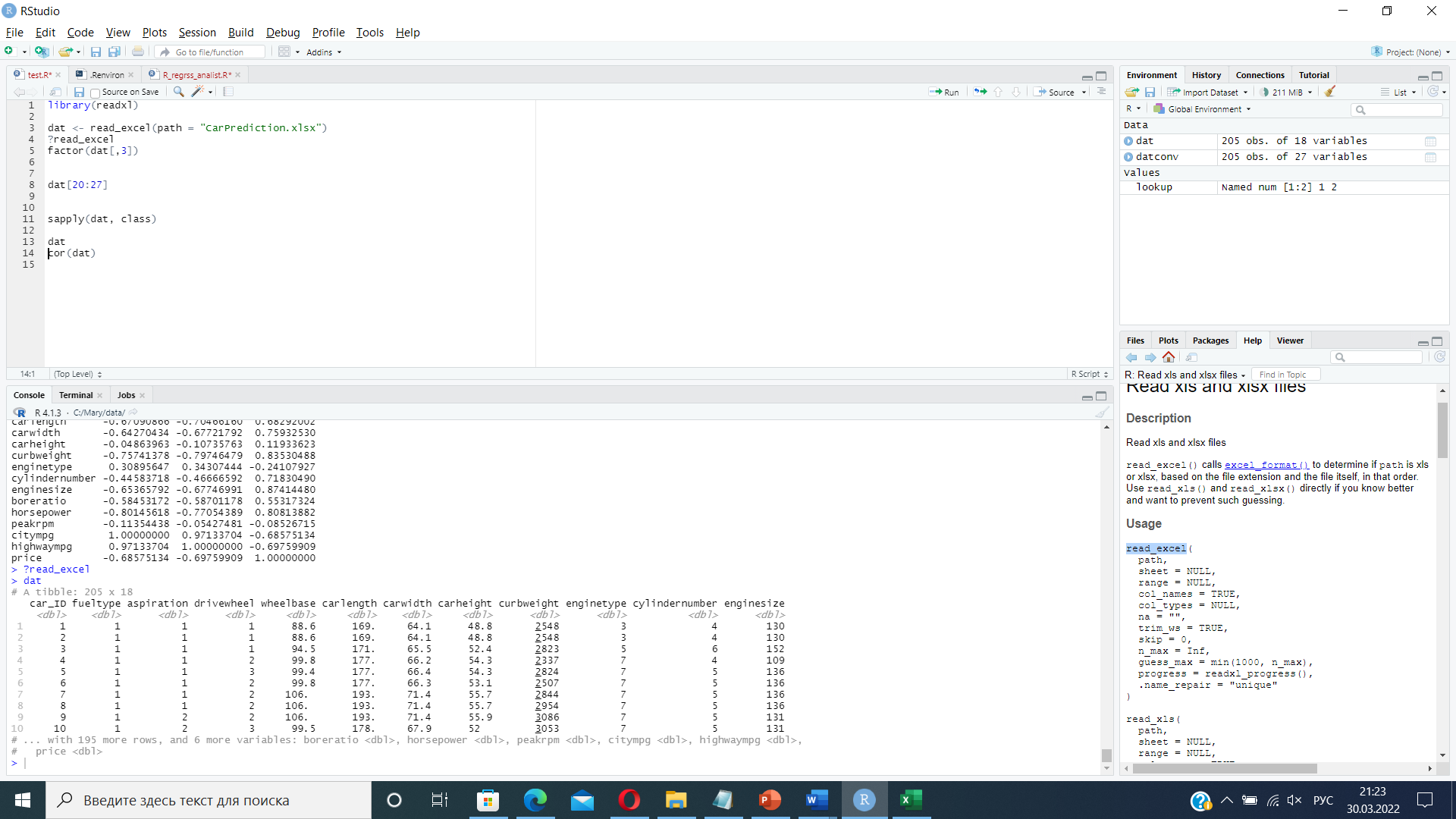
Прочитаем данные из Excel, предварительно указав рабочую директорию:



dat <- read\_excel(path = "CarPrediction.xlsx")

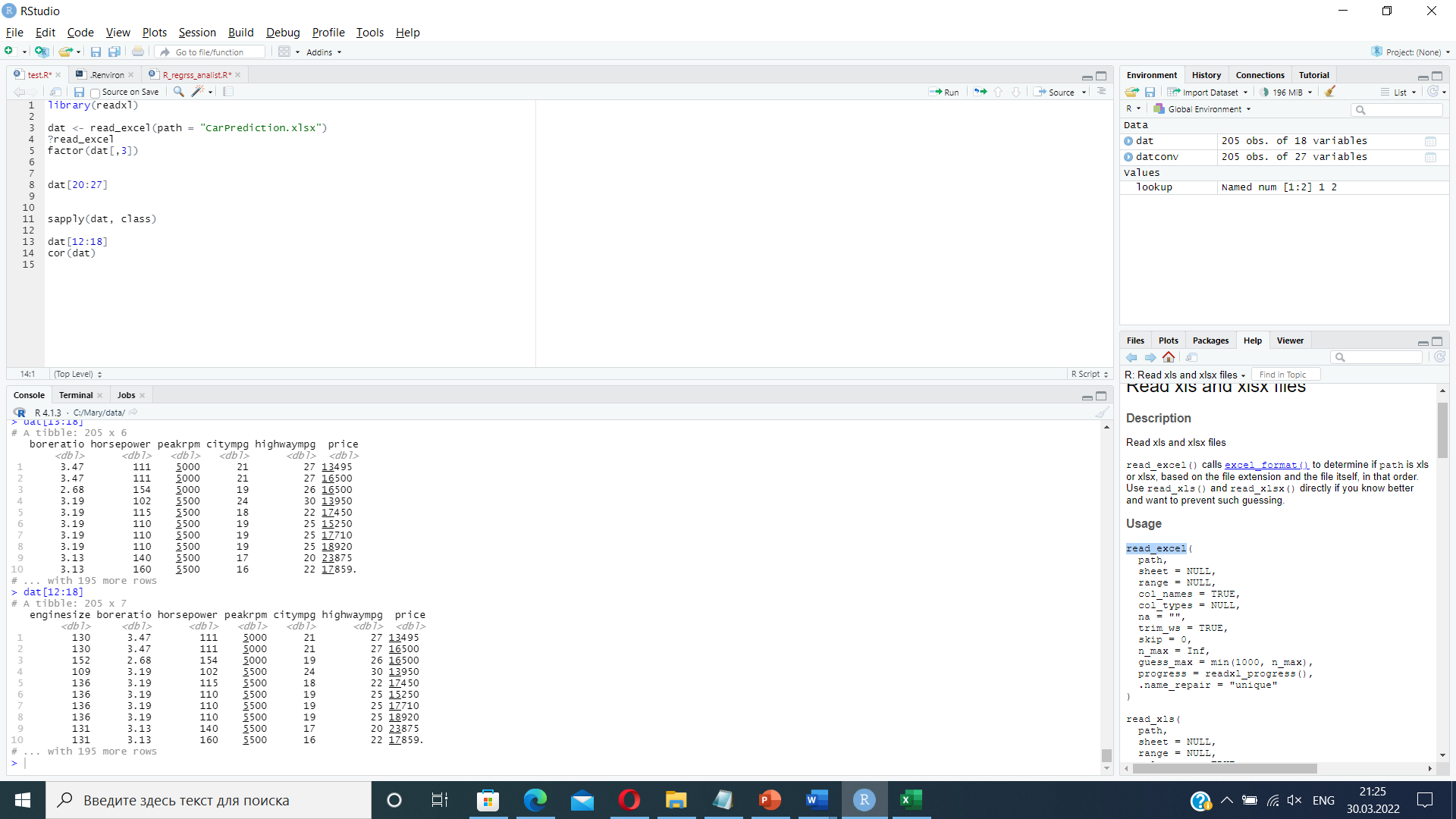
Посмотрим на полученные данные:

dat



Из-за большого числа столбцов не все из них отобразились. Можно указать в квадратных скобках число, указывающее номера столбцов таблицы.

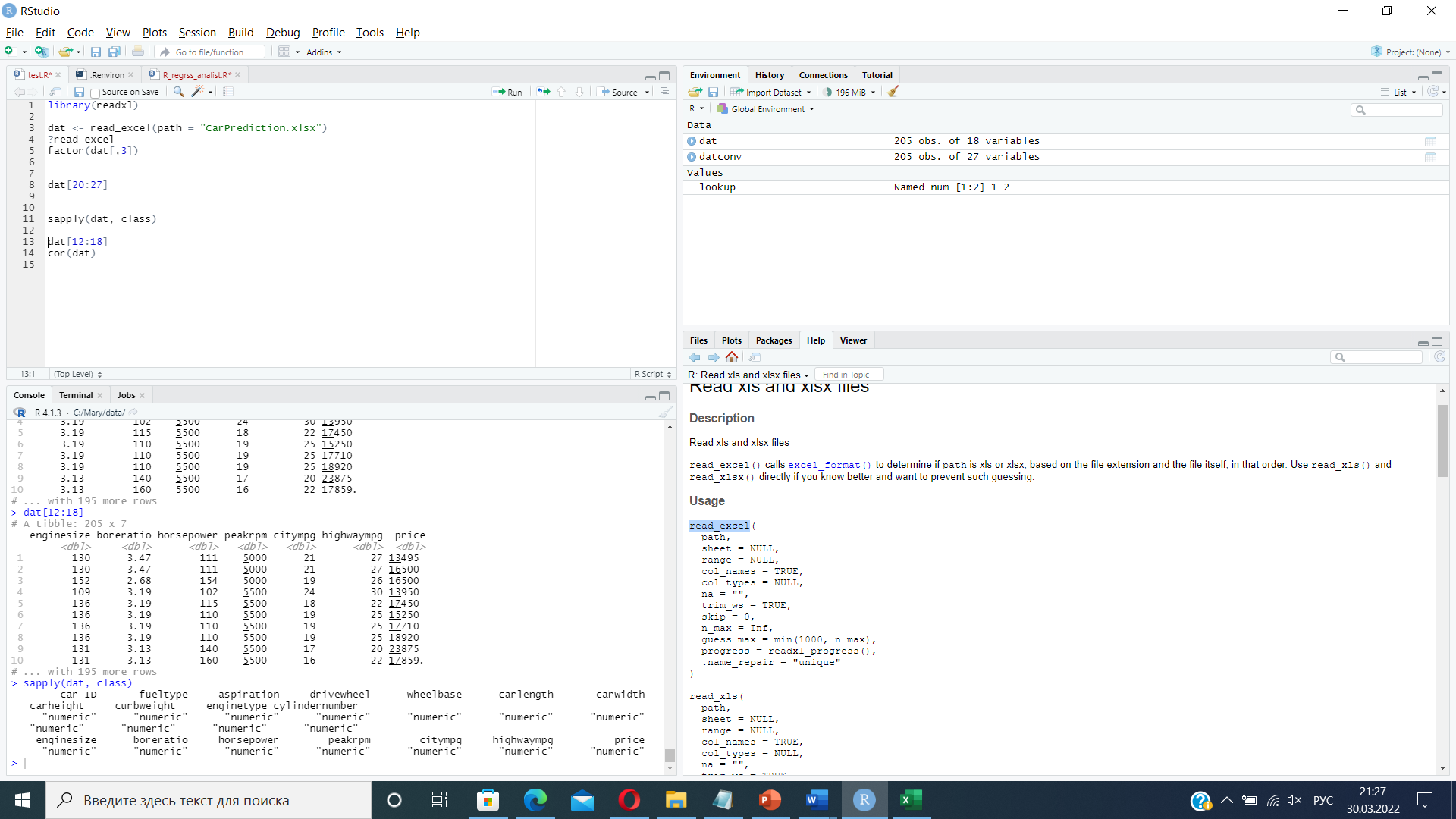
dat[12:18]



Этап 2. Очистка данных.

Проверим типы данных в столбцах

sapply(dat, class)

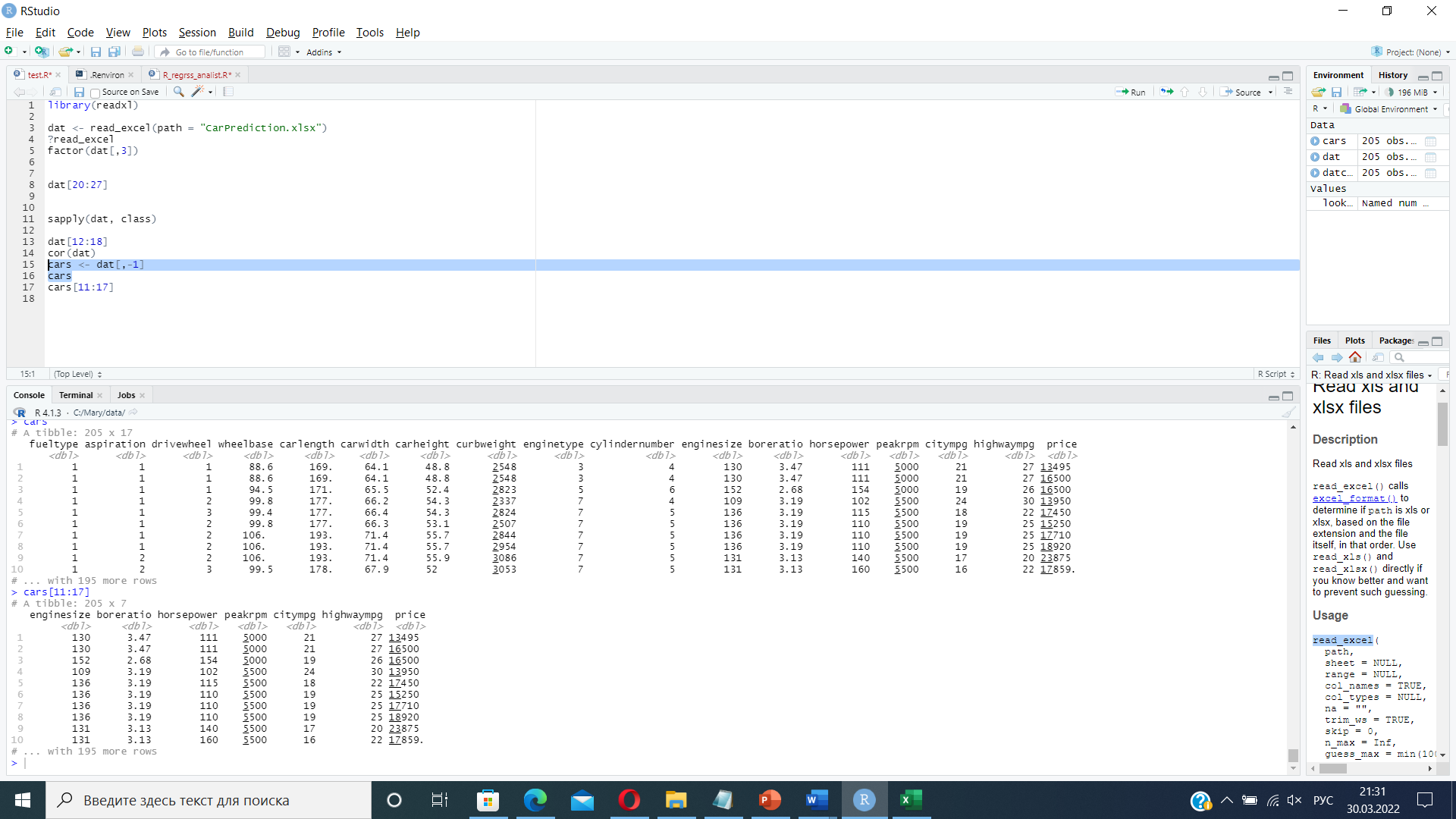


Все столбцы имеют тип numeric – число.

Удалим столбец с идентификатором, поскольку этот фактор не должен влиять на цену авто

cars <- dat[,-1]

cars



Проверим, есть ли «пустые» ячейки в данных (NA):

find\_na <- function(data){

sum =0

for(i in 1:nrow(data)){

for (j in 1:ncol(data))

{

sum<-sum+is.na(cars[i,j])

}

}

print(sum)

}

find\_na(cars)

Результат: 0

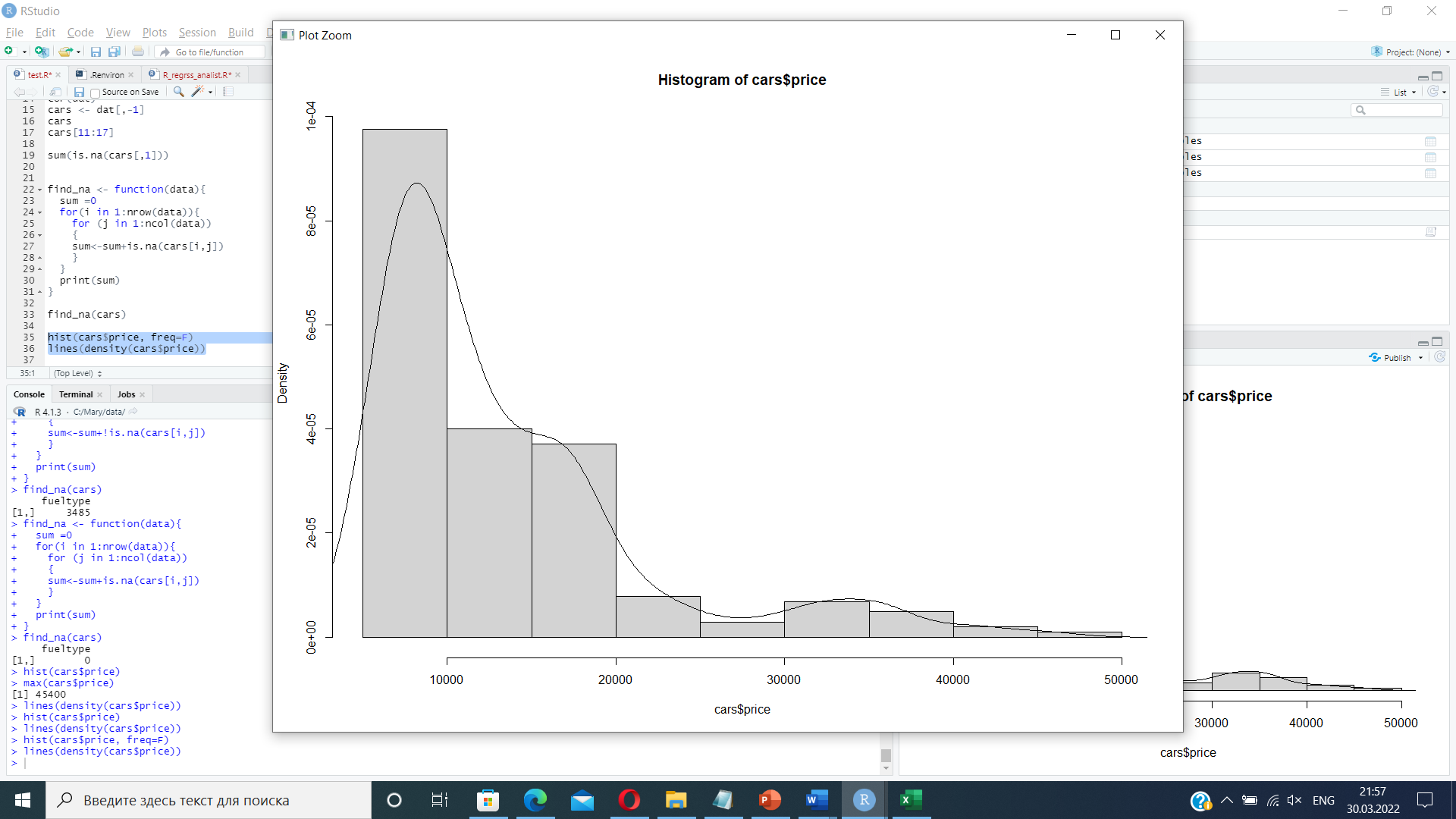
те таких данных нет, все значения представляют собой числа.

Этап 3. Визуализация данных. Выбор независимых переменных

Отобразим распределение цены авто с помощью hist

hist(cars$price, freq=F)

lines(density(cars$price))



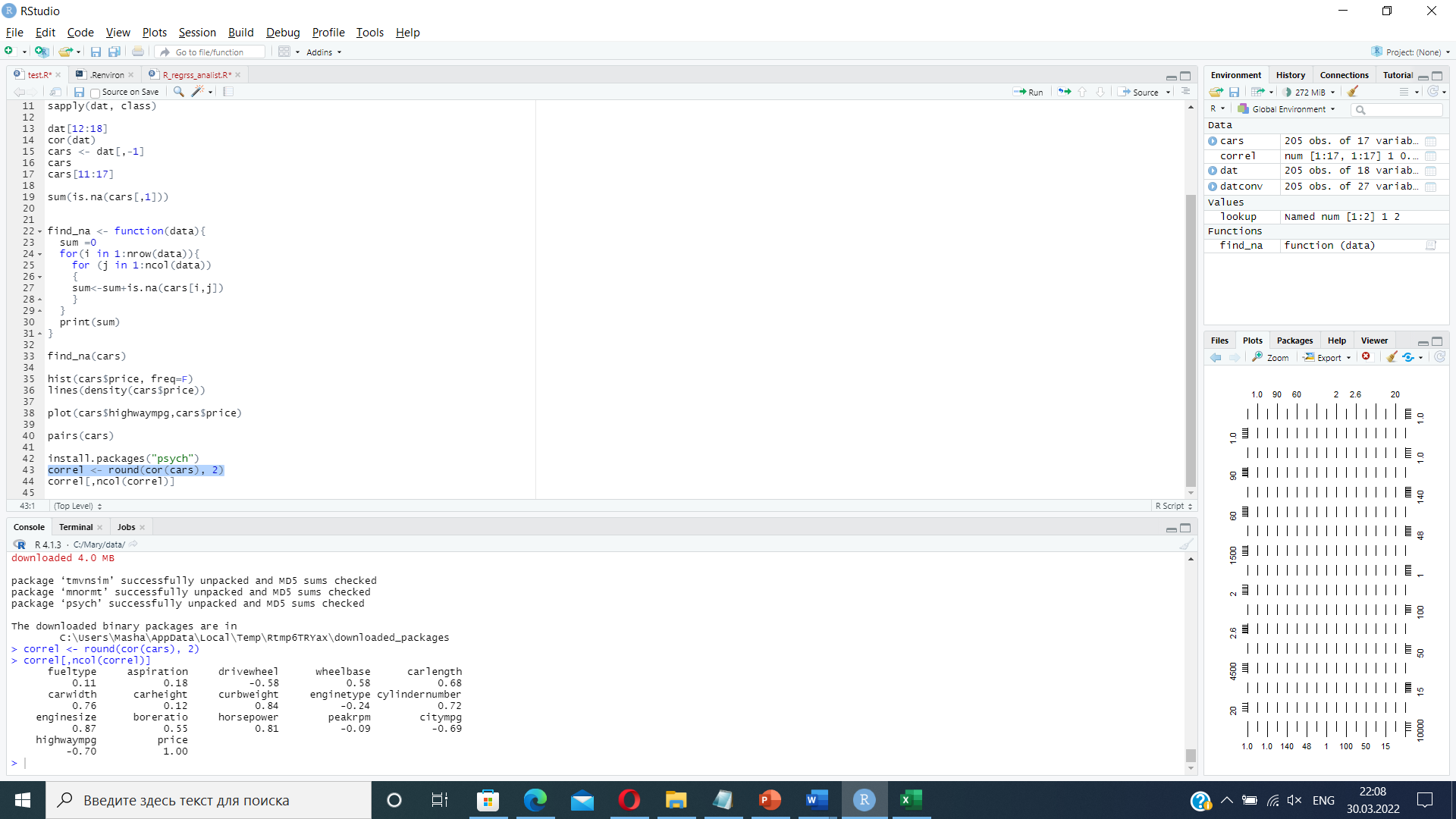
Поскольку в таблице много столбцов, для выявления менее значимых факторов посчитаем матрицу взаимных корреляций всех переменных между собой и округлим результат до двух цифр после запятой. В данном случае нас будет интересовать последний столбец – корреляция цены с другими параметрами.

Для этого воспользуемся пакетом

install.packages("psych")

correl <- round(cor(cars), 2)

correl[,ncol(correl)]

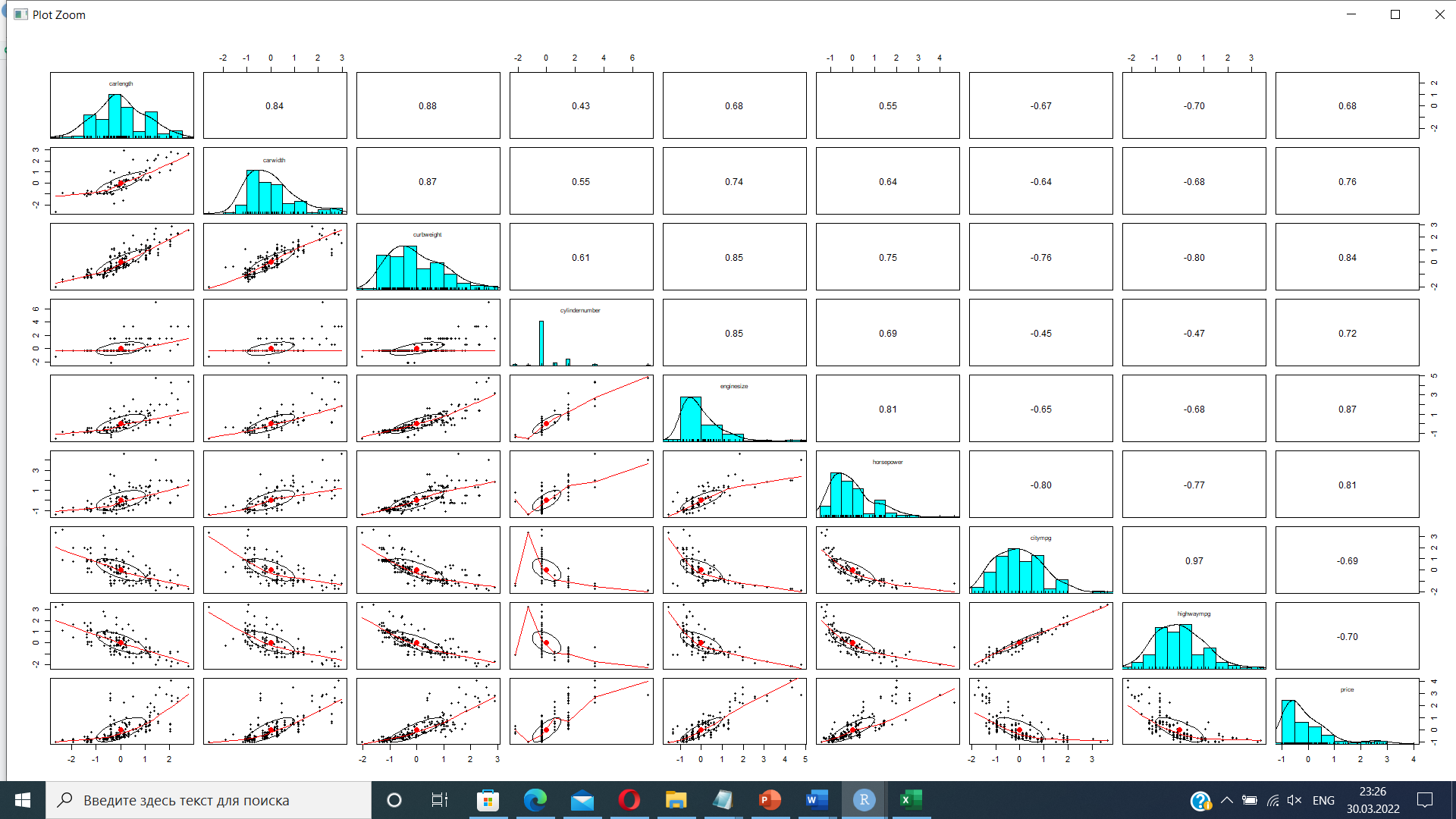


Видно, что столбцы fueltype, aspiration, carheight, enginetype, peakrpm, drivewheel, wheelbase, boreratio по модулю меньше 0.6, что говорит о слабо-умеренной связи с ценой, поэтому также исключим эти факторы из рассмотрения.

cars\_clean <- cars[,c(5:6,8,10:11,13, 15:17)]

Альтернативный способ представления данных - использование функции pairs.panels пакета psych, которая возвращает одновременно и диаграммы распределения данных, и значения коэффициентов корреляции.

psych::pairs.panels(cars\_clean)



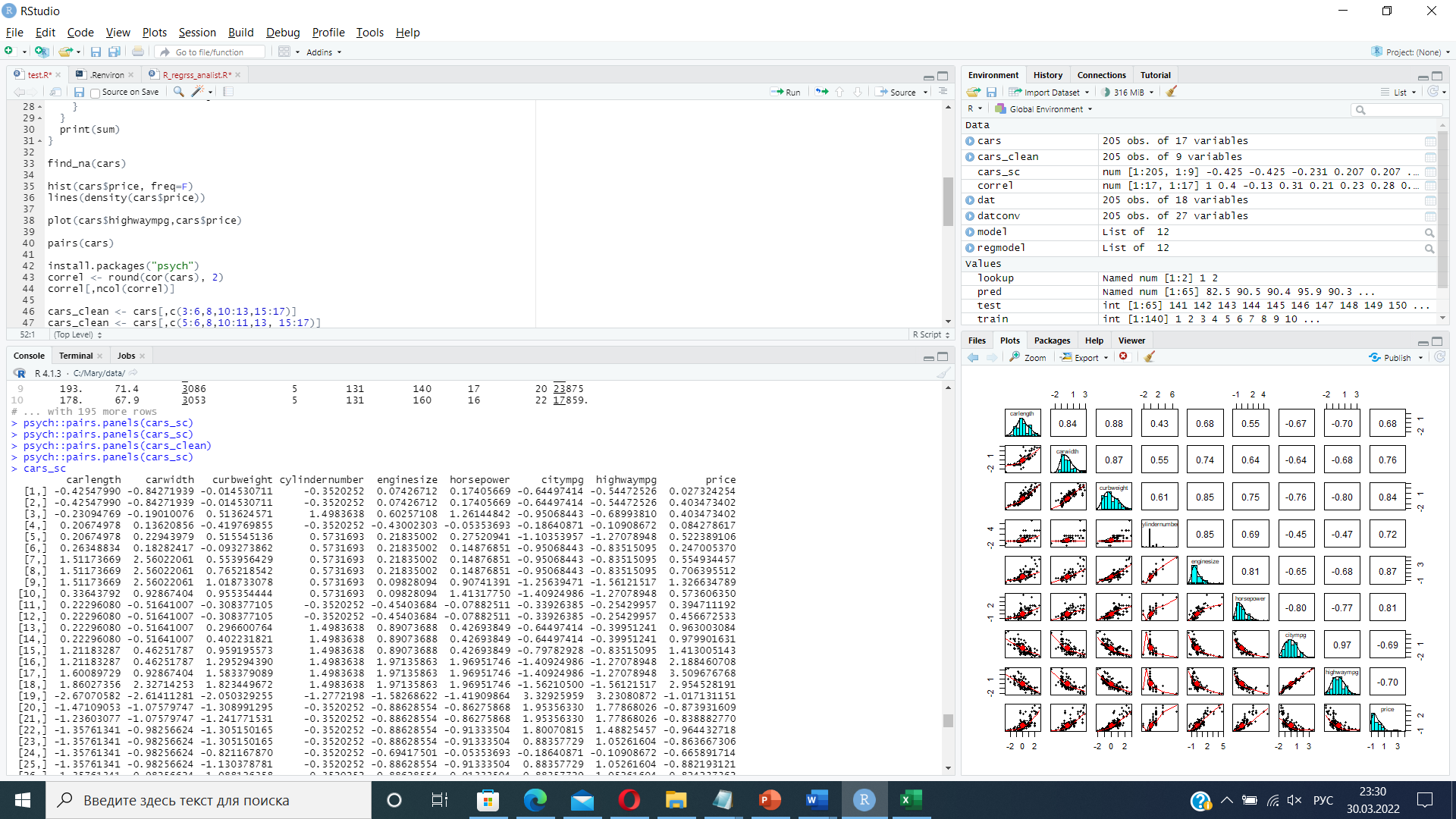
Видно, что все параметры имеют нормальное распределение, что также соответствует требованиям регрессии.

Этап 4. Подготовка данных.

Поскольку некоторые представленные параметры отличаются в 1000 раз (имеют разные единицы измерения), их необходимо стандартизировать.

cars\_sc <- scale(cars\_clean)

cars\_sc



Этап 5. Разделение данных на наборы для обучения и тестирования.

Используем для построения модели около 70% представленных данных, и для тестирования качества модели в дальнейшем оставим около 30%.

train <- 1:140

test <- 141:(nrow(cars\_sc))

Этап 6. Построение линейной модели.

В качестве модели выберем линейную множественную регрессию. Возьмем 6 параметров с наибольшим коэффициентов корреляции с ценой:

model <- lm(price ~ horsepower+enginesize+carwidth+carlength+curbweight+cylindernumber, data = as.data.frame(cars\_sc[train,]))

До знака тильды в функции указывается целевая переменная, после – список выбранных факторов.

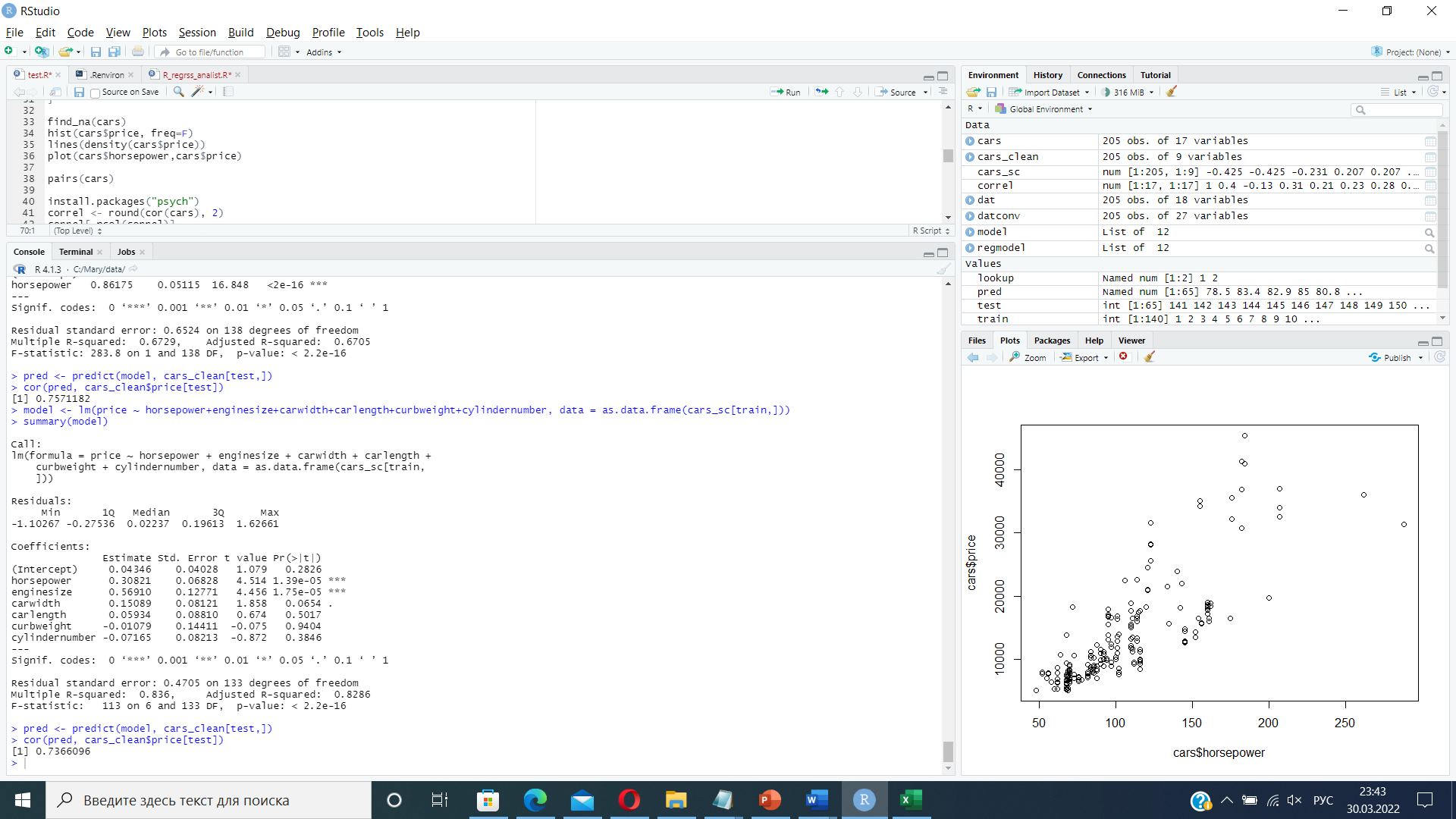
Чтобы посмотреть сведения о линейной аппроксимации используется функция

summary(model)

А также сопоставим качество прогноза с истинным значением

pred <- predict(model, cars\_clean[test,])

cor(pred, cars\_clean$price[test])



Перейдем теперь к расшифровке полученных результатов.

Intercept — точка пересечения прямой с осью координат, те остаточный член.

Estimate – коэффициенты линейной регрессии.

R-squared — коэффициент детерминации; указывает, насколько тесной является связь между факторами регрессии и зависимой переменной. Чем ближе к 1, тем ярче выражена зависимость. В данном случае равен 0.836, что является неплохим результатом.

F-statistic — используется для оценки значимости модели регрессии в целом (чем больше значение параметра, тем лучше).

t value — критерий, основанный на t распределении Стьюдента. Значение параметра в линейной регрессии указывает на значимость фактора, можно считать, что при t>2 фактор является значимым для модели.

p-value — вероятность истинности нуль гипотезы, которая гласит, что независимые переменные не объясняют динамику зависимой переменной. Если значение p-value ниже порогового уровня (0.05), то нуль гипотеза ложная. Чем ниже — тем лучше.

В данном случае видно, что у некоторых переменных t-value значительно ниже 2, поэтому стоит исключить их из рассмотрения также.

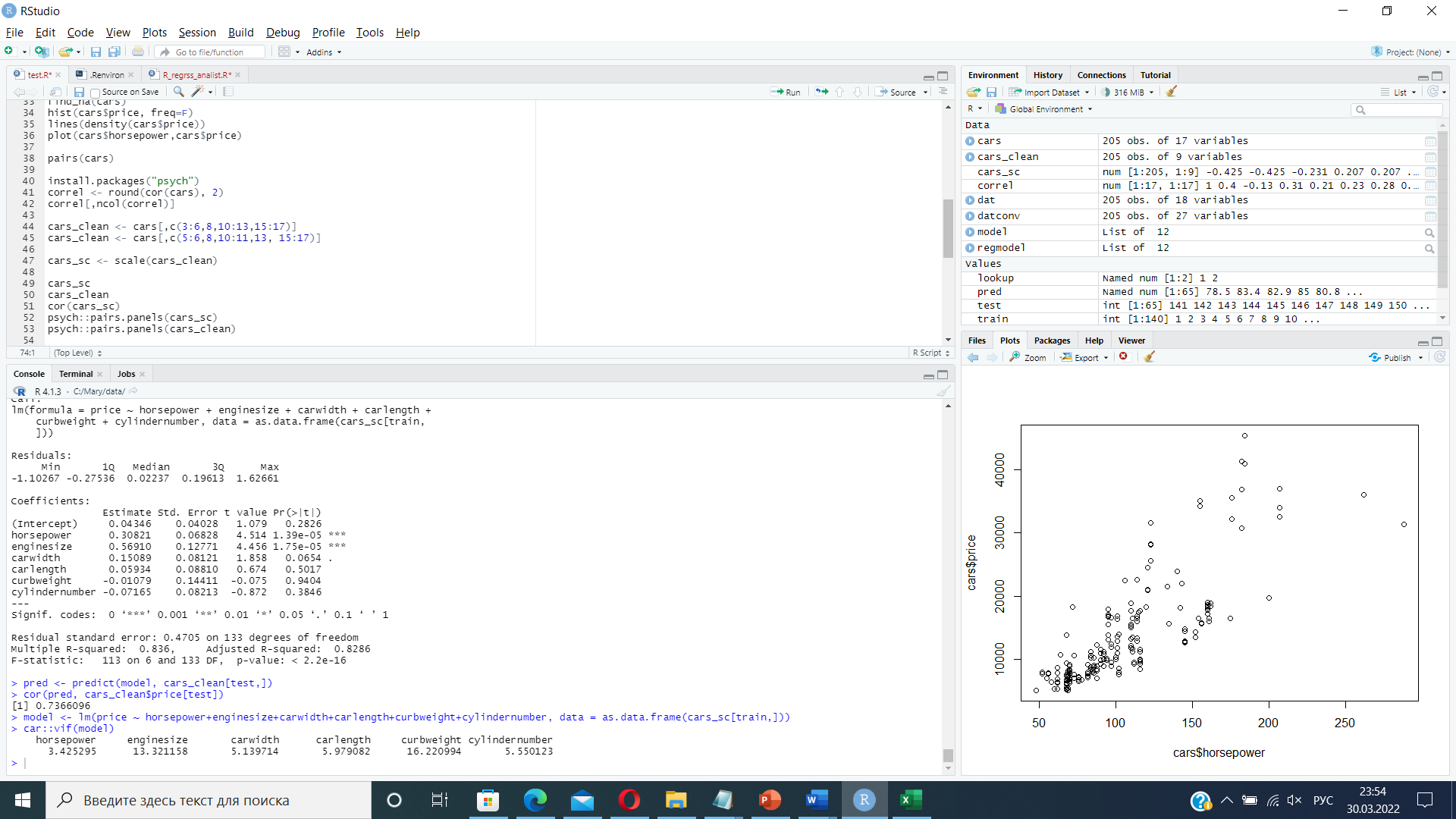
Кроме того, вычислим значение vif (отвечает за мультиколлинеарность, если больше 5-10)

install.packages("usdm")

install.packages("car")

library(car)

car::vif(model)



По результатам видно, что в модели присутствует мультиколлинеарность.

С учетом сказанного выше, исключим из модели незначимые факторы (3 последних).

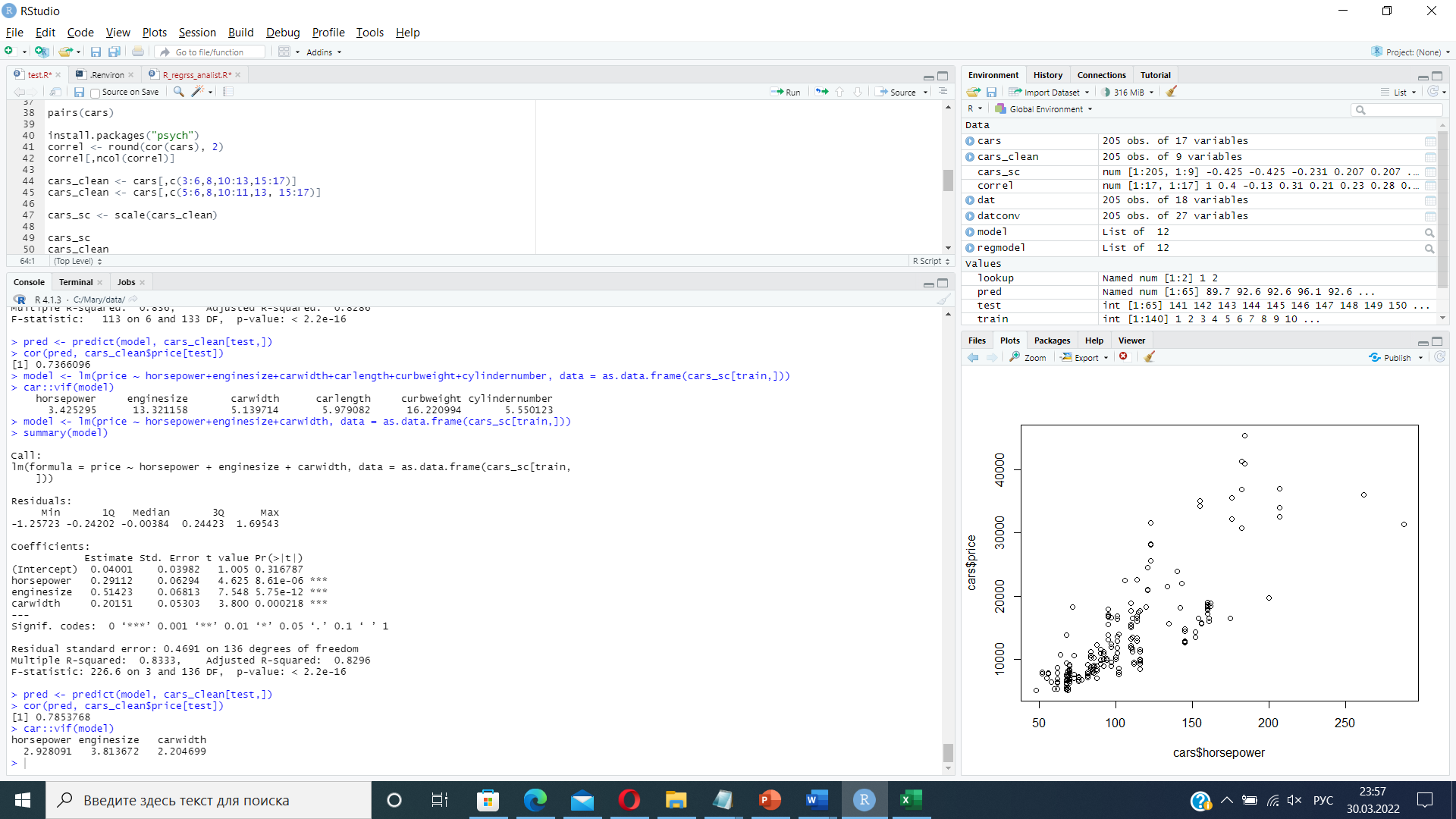
model <- lm(price ~ horsepower+enginesize+carwidth, data = as.data.frame(cars\_sc[train,]))

summary(model)

pred <- predict(model, cars\_clean[test,])

cor(pred, cars\_clean$price[test])

car::vif(model)



Внесенные изменения повлияли на качество модели: исчезла мультиколлинеарность (vif<5), коэффициенты (кроме остаточного члена) являются значимыми (t value > 2), кроме того, качество прогнозирования улучшилось (0.78>0.73).

Остаточный член также можно исключить:

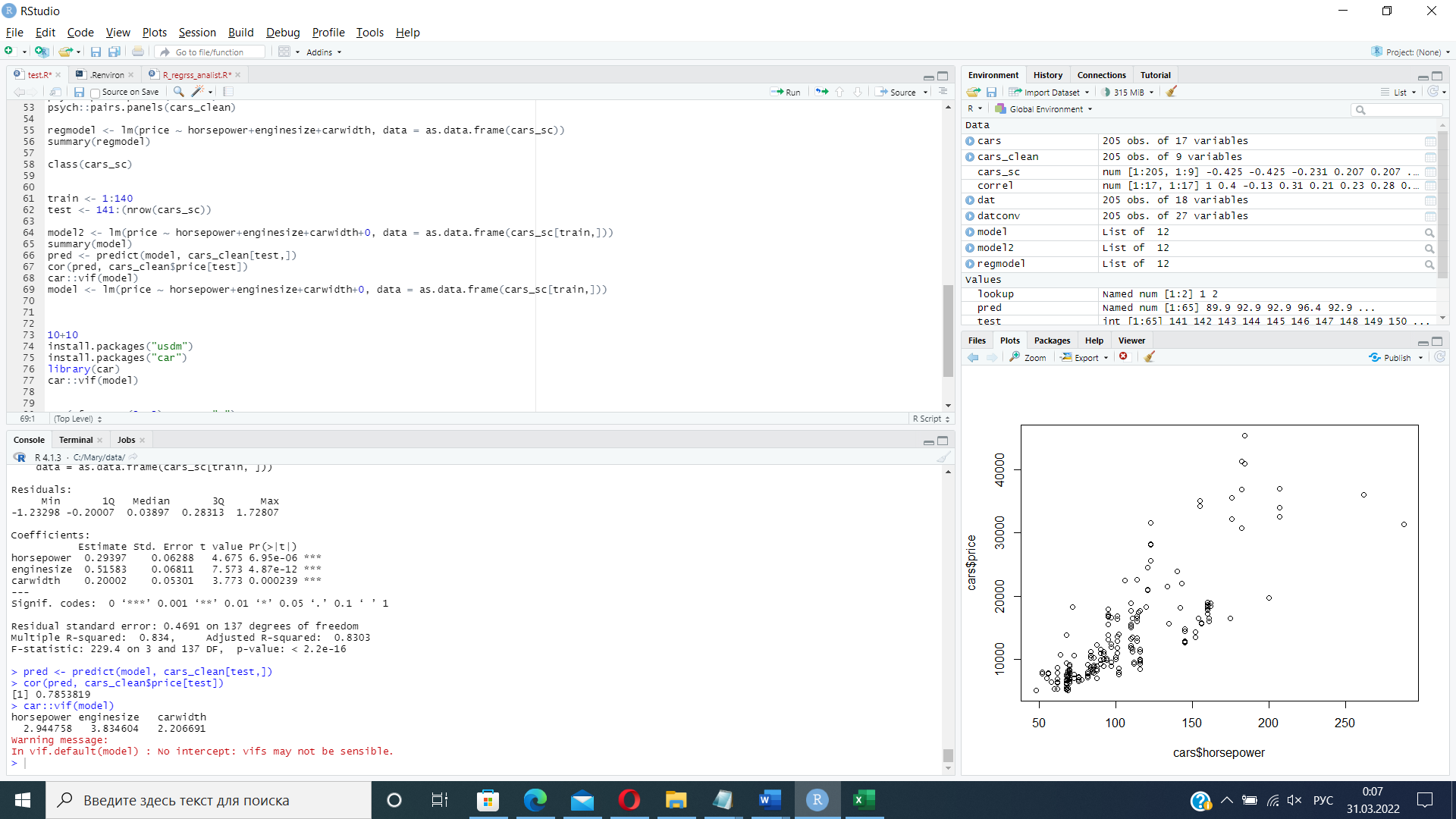
model2 <- lm(price ~ horsepower+enginesize+carwidth+0, data = as.data.frame(cars\_sc[train,]))

summary(model2)

pred <- predict(model2, cars\_clean[test,])

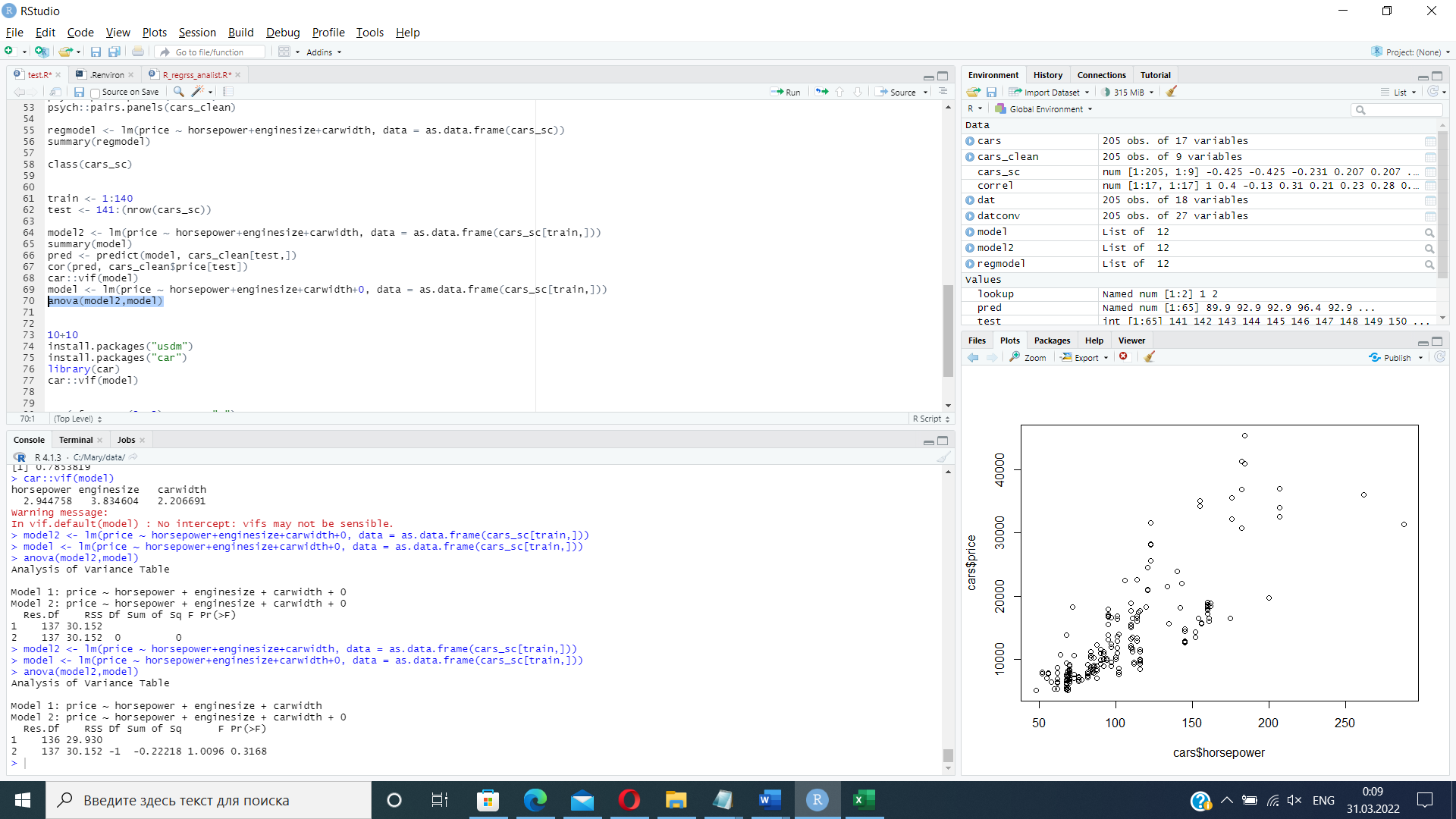
cor(pred, cars\_clean$price[test])

car::vif(model2)



Определим с помощью дисперсионного анализа является ли отличие двух последних моделей значимым или нет.

anova(model2,model)

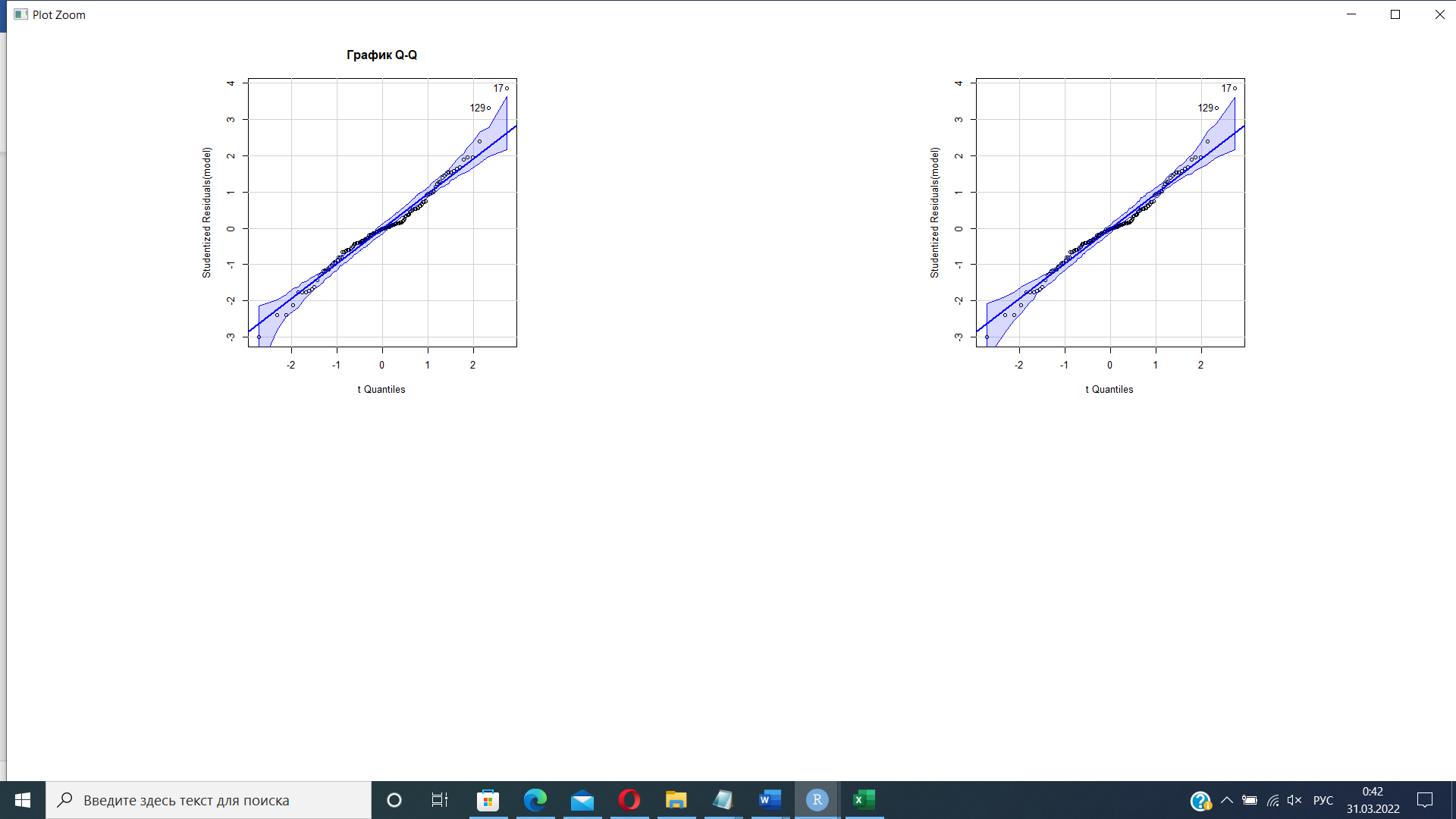


Величина 0.3168>0.05, поэтому можно утверждать, что с вероятностью 95% отличие моделей не значимо и мы в праве выбрать любую модель.

После проверки значимости у полученной регрессионной модели выполняется анализ остатков (разница между прогнозируемым значением и фактическим значением), который должен следовать нормальному распределению.

car::qqPlot(model2, simulate = TRUE)

На графике остаточного QQ точки данных расположены практически по диагональной линии, стремящейся быть прямой, и непосредственно пересекаются диагональю, интуитивно соответствующей нормальному распределению.



Таким образом, полученная регрессионная модель:

price = horsepower\*0.29397 + enginesize\*0.51583 + carwidth\*0.20002,

те цена в большей степени определяется числом лошадиных сил, шириной машины и типом двигателя.

Полученную модель можно оптимизировать, выбирая другие зависимости (нелинейные).

**Список использованных источников:**

1. [Теоретический материал\_регрессионный парн.анализ.pdf (vyatsu.ru)](https://e.vyatsu.ru/pluginfile.php/462626/mod_resource/content/2/%D0%A2%D0%B5%D0%BE%D1%80%D0%B5%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9%20%D0%BC%D0%B0%D1%82%D0%B5%D1%80%D0%B8%D0%B0%D0%BB_%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9%20%D0%BF%D0%B0%D1%80%D0%BD.%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7.pdf)

2. <http://main.isuct.ru/files/publ/PUBL_ALL/167.pdf>

3. [EconometricsWithR.pdf - Яндекс.Документы (yandex.ru)](https://docs.yandex.ru/docs/view?tm=1648581221&tld=ru&lang=ru&name=EconometricsWithR.pdf&text=%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9%20%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7%20%D0%B2%20%D1%81%D1%80%D0%B5%D0%B4%D0%B5%20r&url=https%3A%2F%2Fmeit.mgimo.ru%2Fsites%2Fdefault%2Ffiles%2FEconometricsWithR.pdf&lr=213&mime=pdf&l10n=ru&sign=444a27ee1095a9592d04e2c1b8447e65&keyno=0&serpParams=tm%3D1648581221%26tld%3Dru%26lang%3Dru%26name%3DEconometricsWithR.pdf%26text%3D%25D1%2580%25D0%25B5%25D0%25B3%25D1%2580%25D0%25B5%25D1%2581%25D1%2581%25D0%25B8%25D0%25BE%25D0%25BD%25D0%25BD%25D1%258B%25D0%25B9%2B%25D0%25B0%25D0%25BD%25D0%25B0%25D0%25BB%25D0%25B8%25D0%25B7%2B%25D0%25B2%2B%25D1%2581%25D1%2580%25D0%25B5%25D0%25B4%25D0%25B5%2Br%26url%3Dhttps%253A%2F%2Fmeit.mgimo.ru%2Fsites%2Fdefault%2Ffiles%2FEconometricsWithR.pdf%26lr%3D213%26mime%3Dpdf%26l10n%3Dru%26sign%3D444a27ee1095a9592d04e2c1b8447e65%26keyno%3D0)

4. <https://habr.com/ru/post/207750/>

5. <http://qsar4u.com/files/rintro/03.html>

6. <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction>