ЛАБОРАТОРНАЯ РАБОТА №1

Визуализация данных с помощью ggplot2 (функция qplot)

Цель работы: познакомиться с основными элементами графической грамматики графики пакета ggplot2.

ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

Построение графиков с использованием пакета ggplot2 представляет набор шагов, каждый из которых можно назвать грамматической единицей. Комбинируя набор шагов, мы получаем необходимый график.

Принципы грамматики графики в ggplot2

Основные принципы грамматики графики включают следующие шаги:

Aesthetic attributes – выбор данных из набора, которые будут использоваться для построения итогового изображения. Выбираются переменные для отображения на осях (оси) графика и переменные для атрибутов изменения объектов, располагаемых на самом графике.

Geometric object – выбор варианта того, как именно данные будут представлены на графике (точки, линии, столбцы и др.).

Statistical transformations – выбор статистических трансформаций (регрессионная прямая, сглаживание).

Scales – работа с масштабом и отображением данных.

Coordinates – выбор системы координат.

Faceting – группировка данных.

В пакете ggplot2 заложена два принципиально отличных по сложности подхода для построения графиков. Первый, наиболее каноничный, использует функцию ggplot и соответствующие элементы грамматики графики. Существует и упрощенный – второй вариант построения графиков с помощью функции qplot.

Работа с функцией qplot

qplot в большей степени походит на базовую графику пакета R и как бы «пытается угадать» то, что пользователь планирует визуализировать. «Угадывание» происходит на основании количества и типа передаваемых переменных. Для визуализации всего сказанного будем использовать набор данных diamonds, содержащийся в пакете ggplot2.

diamonds – набор данных, содержащий сведения о 53940 бриллиантах. Используется 10 переменных: price, carat, cut, color, clarity, x, y, z, depth и table (рисунок 1).

```
# A tibble: 53,940 x 10
   carat cut       color clarity depth table price     x     y     z
   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
 1 0.23  Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
 2 0.21  Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
 3 0.23  Good      E     VS1      56.9    65   327  4.05  4.07  2.31
 4 0.290 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
 5 0.31  Good      J     SI2      63.3    58   335  4.34  4.35  2.75
 6 0.24  Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
 7 0.24  Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
 8 0.26  Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
 9 0.22  Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
10 0.23  Very Good H     VS1      59.4    61   338  4     4.05  2.39
# … with 53,930 more rows
```

Рисунок 1 – Структура набора данных diamonds пакета ggplot2

В зависимости от того, сколько и какие переменные передаются в функцию qplot, она будет использовать разные геометрии для визуализации. Ниже представлены три варианта передачи разных переменных (рисунок 2)

```
qplot(x = price, data = diamonds)
qplot(x = price, y = carat, data = diamonds)
qplot(x = cut, y = carat, data = diamonds)
```

Рисунок 2 – Примеры вызова функции qplot

На рисунке 3 показаны визуализации трех разных вариантов передачи переменных.

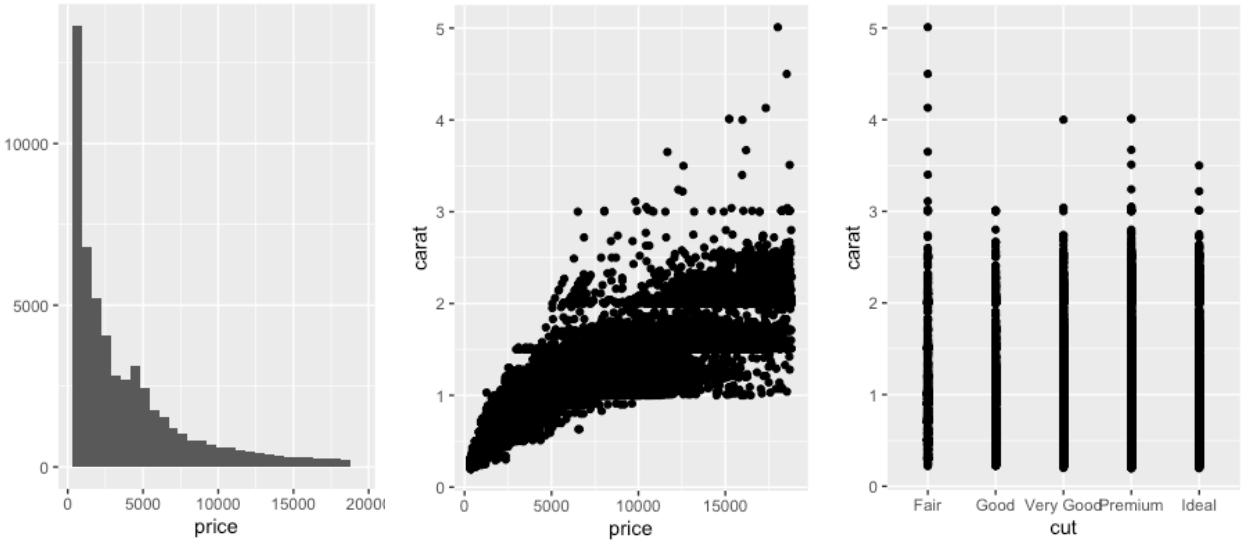


Рисунок 3 – Примеры построения графиков функцией qplot

1) При передаче одной переменной целочисленного типа функция qplot строит гистограмму частот.

2) При передаче одной целочисленной и одной вещественной переменной – диаграмму рассеивания.

3) При передаче одной переменной факторного типа и одной переменной вещественного типа – точечную диаграмму.

Код построения графика можно сохранить в переменную и повторно использовать там, где это необходимо.

Некоторые аргументы функции qplot:

x – переменная x (ось абсцисс);

y – переменная y (ось ординат);

shape – формат точек, принимает или переменную факторного типа или использует запись типа I(#), где вместо # число от 0 до 25 или от 32 до 127;

color – цвет точек принимает количественную или факторную переменную типа или использует наименование цвета в кавычках;

fill – заливка фигур (аналогично color);

size – размер точек принимает или переменную факторного типа или использует запись типа I(#), где вместо # вещественное не отрицательное число;

alpha – прозрачность принимает количественную или факторную переменную или использует запись типа I(#), где вместо # вещественное число в интервале от 0 до 1;

xlim – задание ограничений по оси абсцисс, задается в виде вектора с двумя значениями;

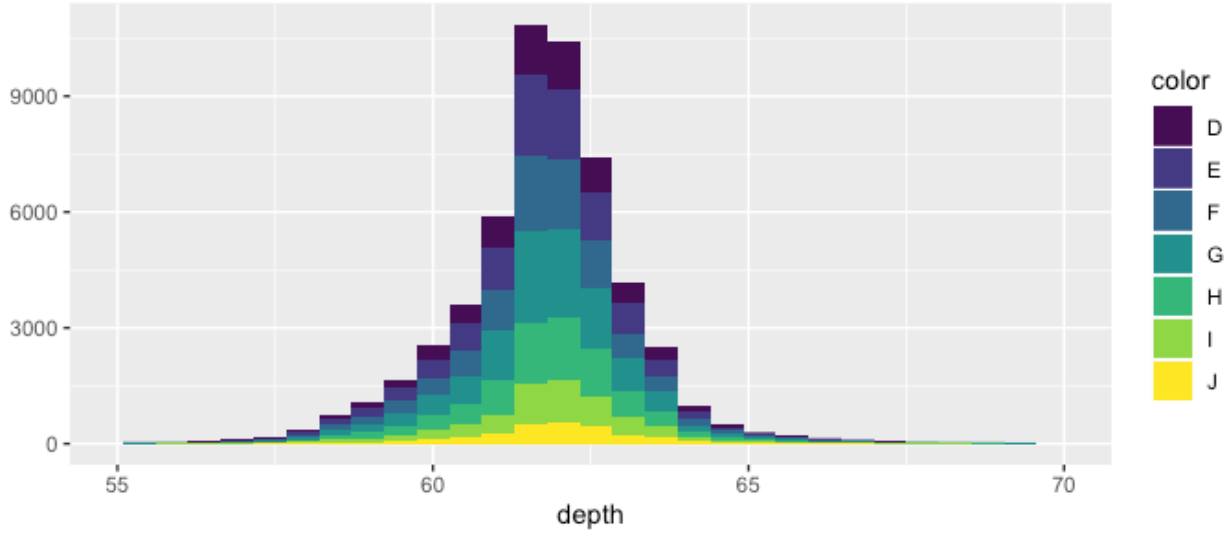
ylim – задание ограничений по оси ординат, задается в виде вектора с двумя значениями;

data – набор данных.

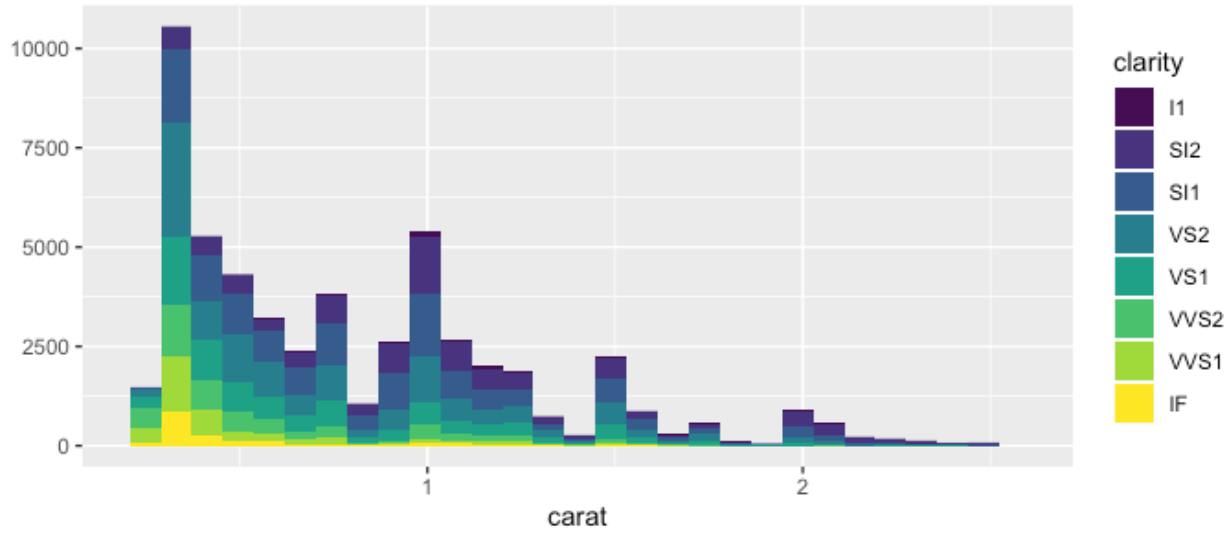
ПРАКТИЧЕСКАЯ ЧАСТЬ

Задание 1 – С помощью функции qplot напишите программный код, строящий следующие графики по набору данных diamonds.

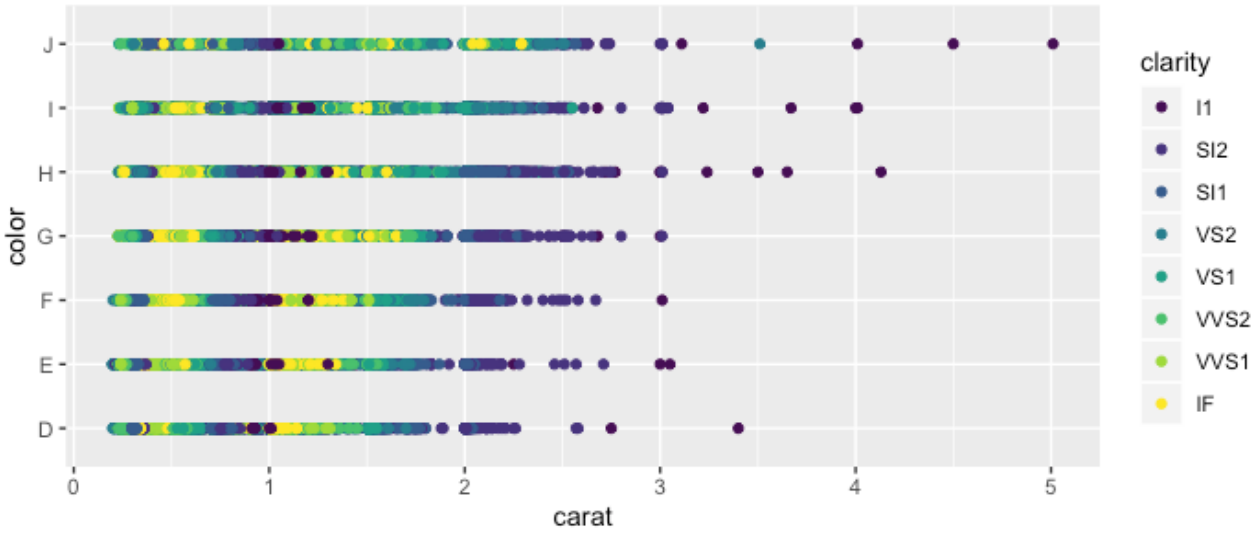
Вариант – 1



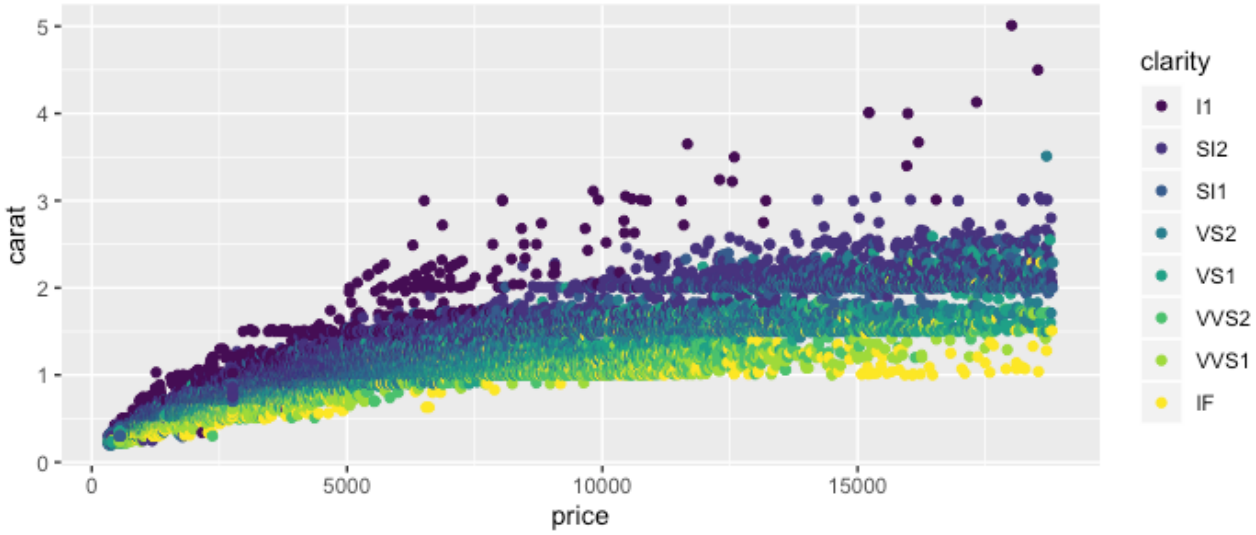
Вариант – 2



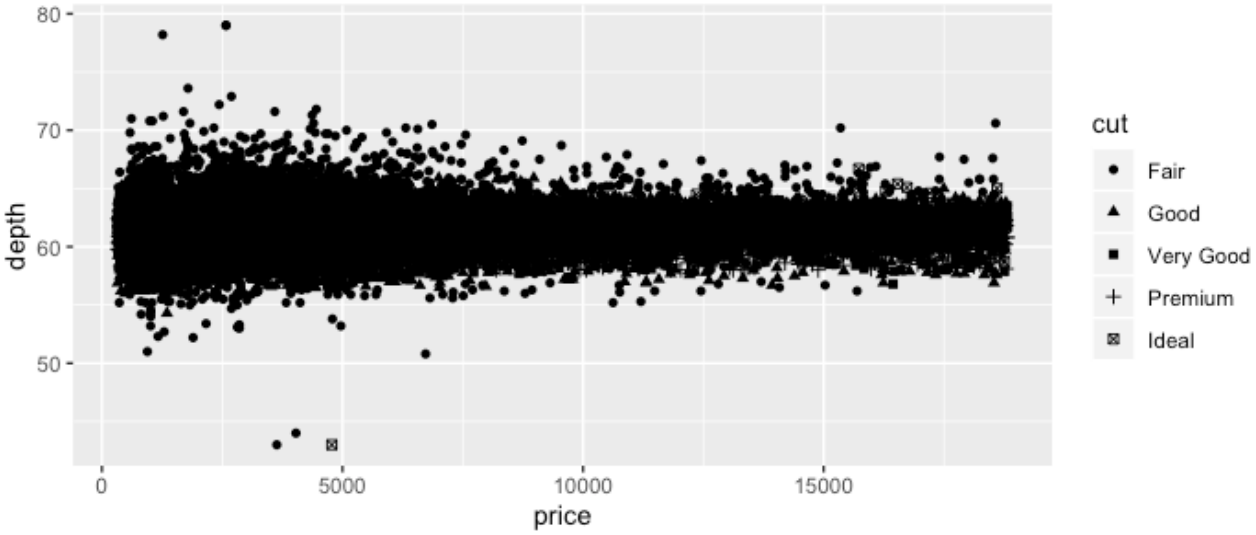
Вариант – 3



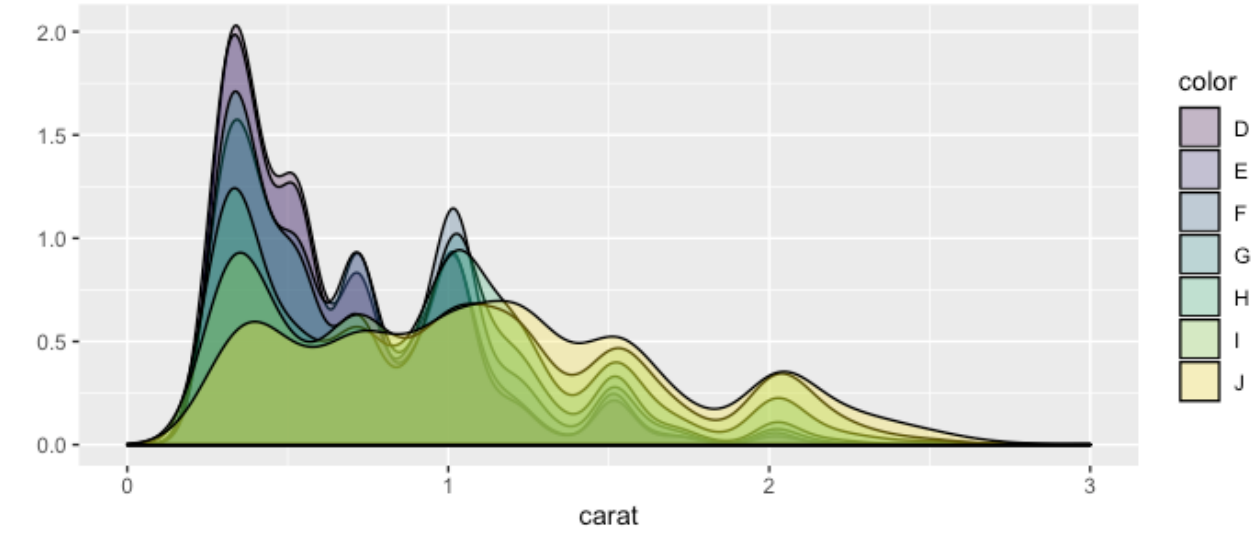
Вариант – 4



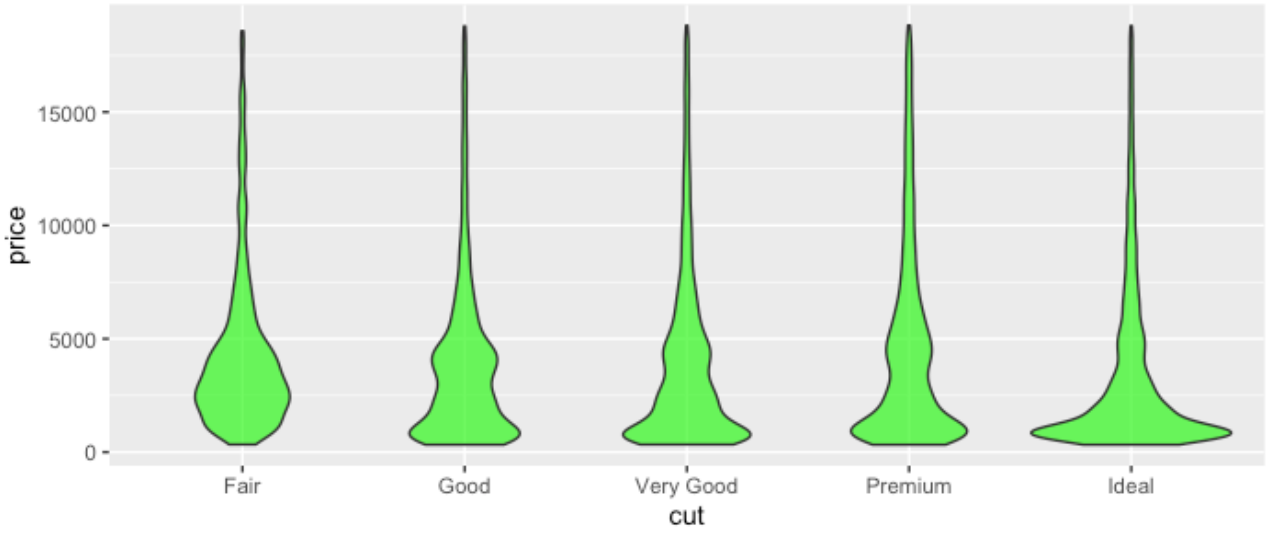
Вариант – 5



Вариант – 6

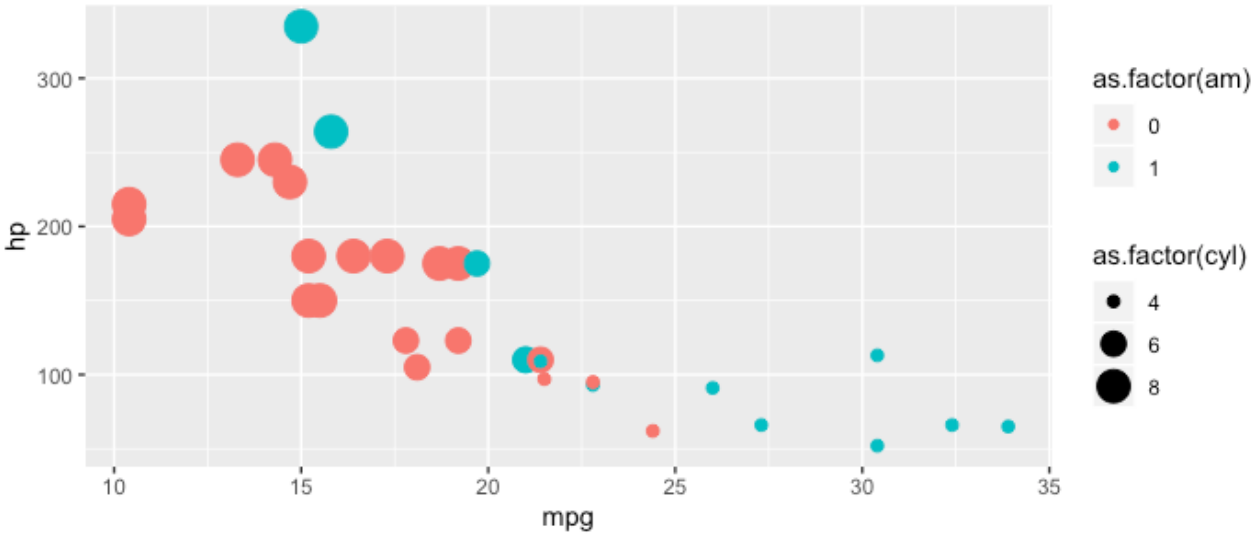


Вариант – 7

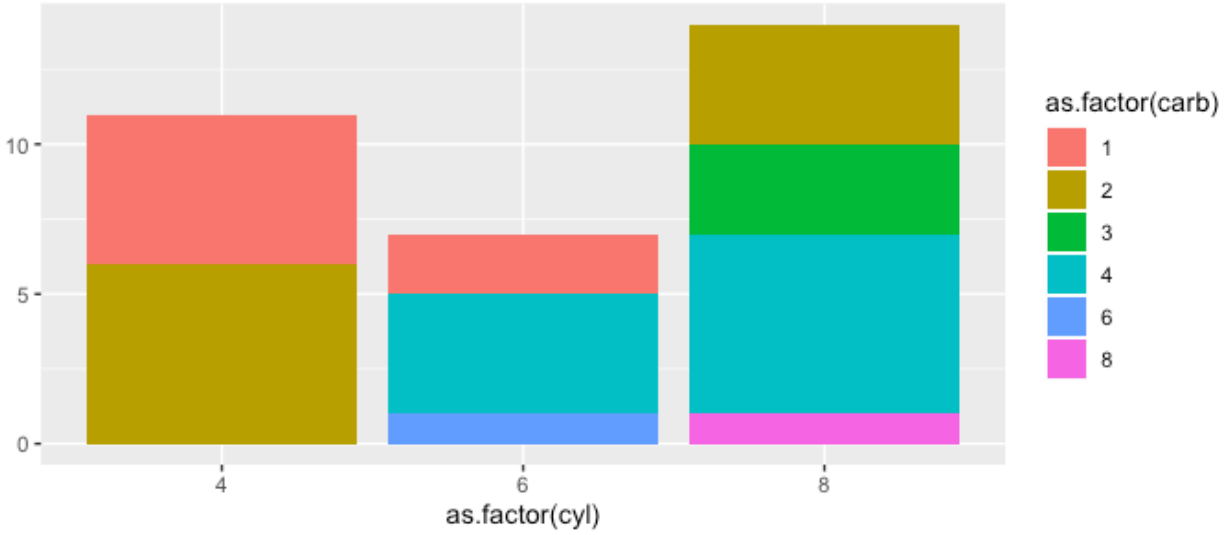


Задание 2 – С помощью функции qplot напишите программный код, строящий следующие графики по набору данных mtcars. При построении графиков потребуется использование функции as.factor()

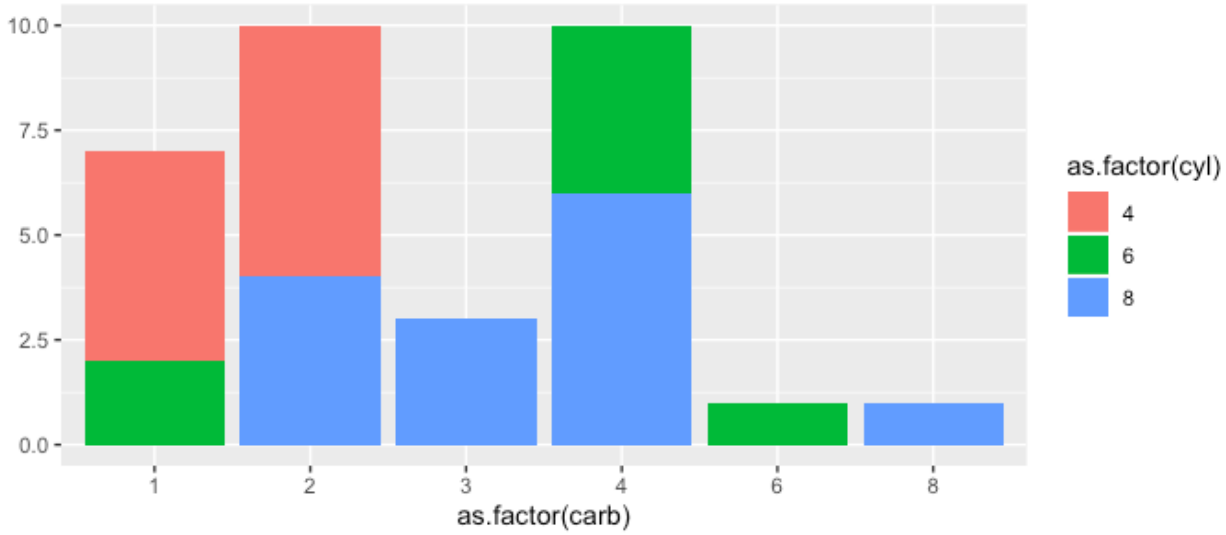
Вариант – 1



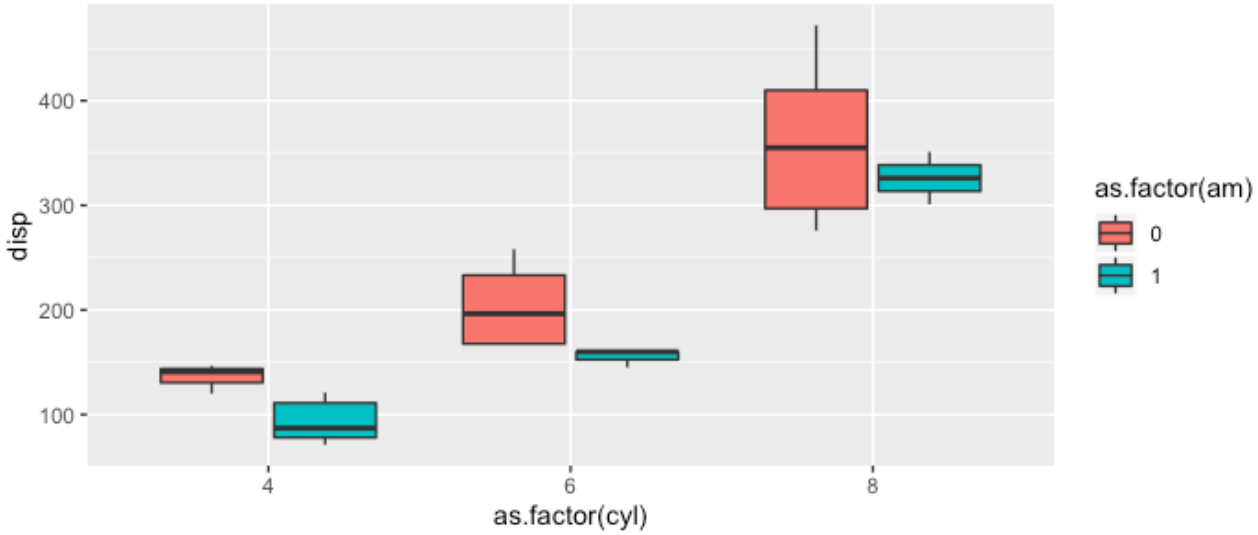
Вариант – 2



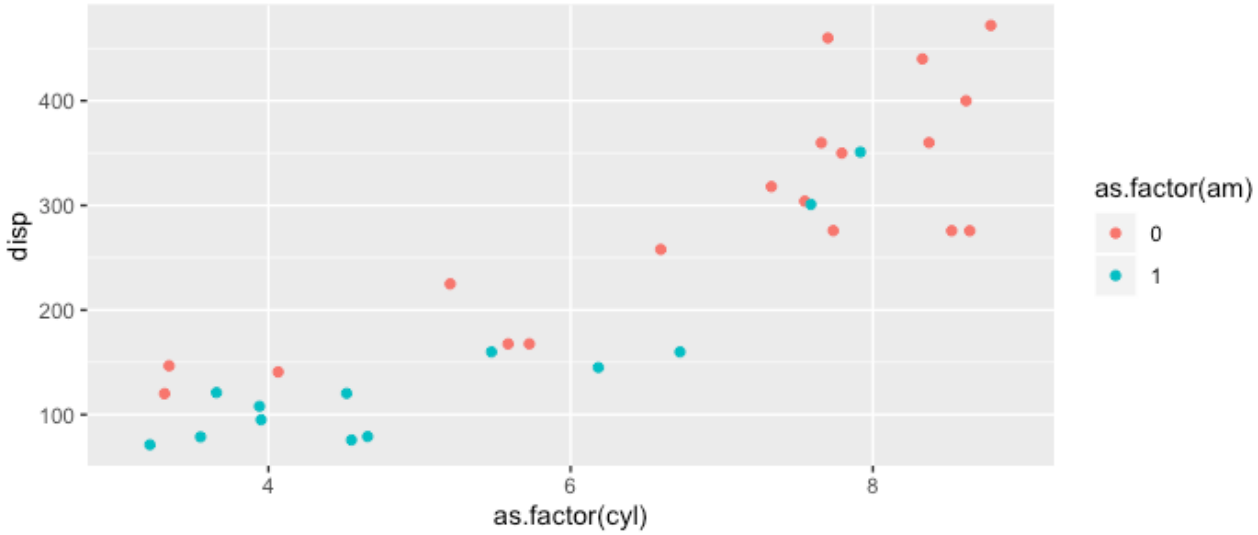
Вариант – 3



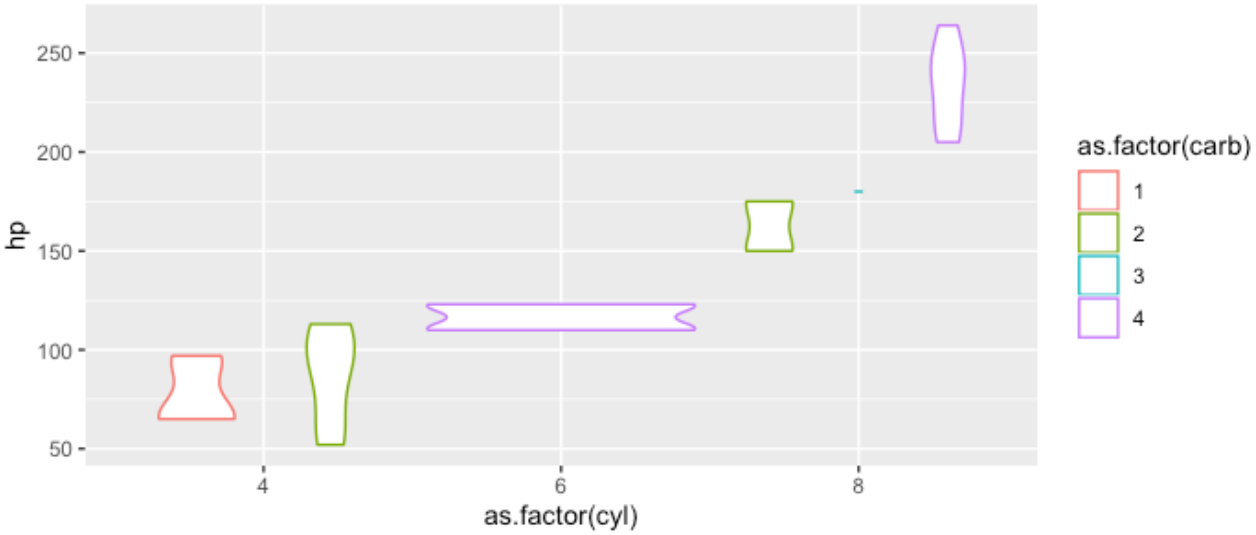
Вариант – 4



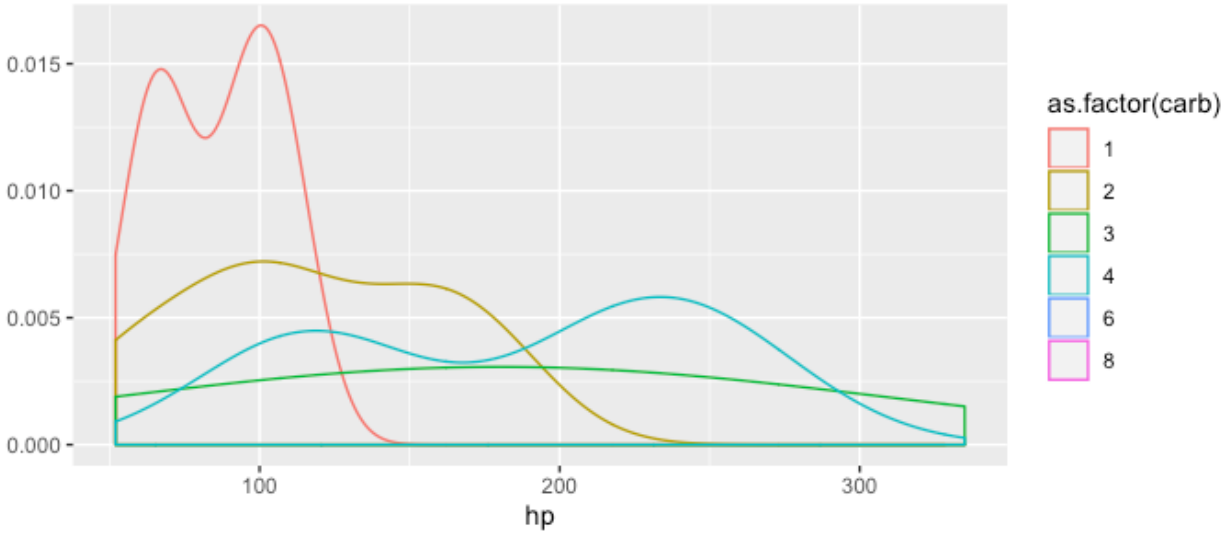
Вариант – 5



Вариант – 6



Вариант – 7



Задание 3 – Постройте два произвольных графика, показав умение использовать различные настройки для набора данных по заданному варианту. Дайте описание набору данных и то, что показывает изображенный график.

№	Набор данных	№	Набор данных	№	Набор данных	№	Набор данных
1	CO2	8	Puromycin	15	iris	22	swiss
2	ChickWeight	9	Seatbelts	16	longley	23	trees
3	DNase	10	Theoph	17	mtcars	24	economics
4	LifeCycleSavings	11	ToothGrowth	18	quakes	25	faithfuld
5	Loblolly	12	USArrests	19	rock	26	midwest
6	Orange	13	freeny	20	stack.x	27	mpg
7	OrchardSprays	14	infert	21	stackloss	28	txhousing

Контрольные вопросы:

1. Для чего применяется функция I()?
2. В чем отличия функции qplot() от ggplot?
3. Какие аргументы используются для настройки изображения?
4. Какой диапазон у аргумента alpha?
5. Что означает запись типа geom = "density"?
6. Для чего применяется конструкция as.factor()?
7. Как произвести заливку элементов зеленым цветом?
8. Каков синтаксис аргумента xlim?