Домашнее задание по базовой визуализации данных в R 1.

Информационная часть

Дата выдачи: 15.09.2024
Дата дедлайна: 28.09.2024 23:59 по Мск (GMT +3)
Максимальный балл без учета дополнительных заданий: 13
Максимально возможный балл: 14
Балл для получения зачета: 10, при этом в каждом блоке должно быть получено не менее 2.5 баллов.

В задании вам предстоит продолжить работать с датасетом hogwarts. Данные и заполненный rmd-файл с прошлого занятия можно найти в папке на гугл-диске.

Полученный rmd и результат knit в формате html загрузите на свой гитхаб. Чтобы сдать задание, приложите ссылку на гитхаб в гугл-классы. До времени дедлайна вы можете вносить любые изменения и дополнения, даже после того, как отправите ссылку. Задание сдается в формате зачет/незачет. Чтобы набрать зачет необходимо набрать как минимум 80% от максимального числа баллов без учета дополнительных (10 баллов из 13-ти) и минимум 2.5 балла из каждого блока.

Некоторые задания требуют обращения к справке и документации по упомянутым на лекциям функциям. Для двух заданий можно получить дополнительные 0.5 балла – это отмечено в тексте. Эти задания потребуют использования подходов, не затронутых на лекции.

Последний блок заданий относится к функциям ggplot2, не упомянутым на лекции. По ним прилагается дополнительная справочная информация. Если у вас возникают трудности технического, коммуникативного или иного характера – можете писать в чат курса в телеграме.

Задания

Столбчатые диаграммы

1. Постройте барплот (столбчатую диаграмму), отражающую распределение числа студентов по курсу обучения. Примените любую из встроенных тем ggplot. Раскрасьте столбики любым понравившимся вам цветом (можно использовать как словесные обозначения, так и гекскоды). Добавьте цвет контура столбиков. (1 б).
2. Создайте новый барплот, отражающий распределение числа студентов по факультету. Добавьте на график вторую факторную переменную – происхождение (bloodStatus). Модифицируйте при помощи аргумента position графика так, чтобы каждый столбец показывал распределение факультета по чистоте крови в долях. Примените произвольную тему. Запишите текстом в rmd-документе, какой вывод можно сделать из графика? (1 б).
3. Модифицируйте датасет таким образом, чтобы в нем остались только чистокровные (pure-blood) и маглорожденные студенты (muggle-born). Создайте на основе этих данных график из пункта 2. Добавьте горизонтальную пунктирную линию произвольного цвета на уровне 50%. Дайте осям название на русском языке (1б). *Дополнительно: переименуйте на русский язык категории легенды pure-blood и muggle-born* (0.5 б).

Боксплоты

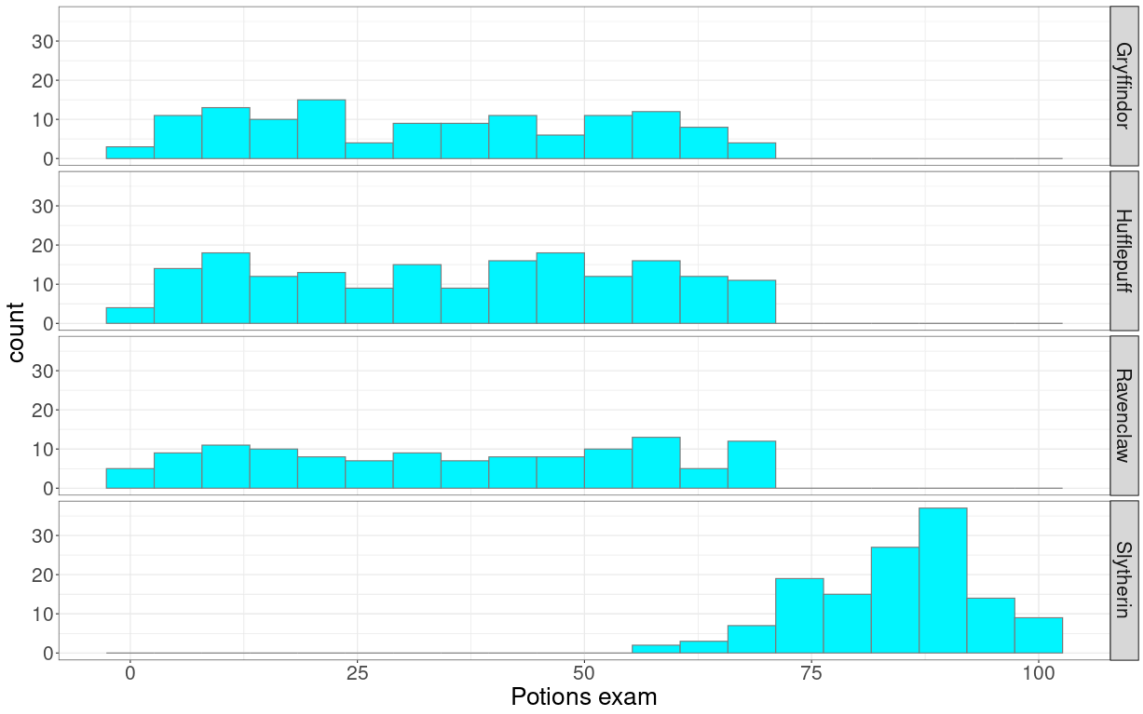
1. Отобразите распределение баллов, заработанных студентами на 3-й неделе обучения, по факультетам. Отсортируйте факультеты в порядке убывания медианного балла за 3-ю неделю (мы не останавливались на этом в лекции, но упомянутая в ней функция по умолчанию сортирует именно по медиане, так что в этом случае дополнительных аргументов передавать не следует). (1 б.)
2. Добавьте отображение разными цветами для происхождения студентов (bloodStatus). Добавьте на боксплот вырезку (`notch`). Настройте для данного чанка размер изображения 14:14 дюймов. Приведите названия осей к корректному виду. (1 б.)
3. Добавьте на график джиттер-плот. Удалите отображение выбросов у боксплота. Видоизмените по своему вкусу толщину линий и ширину боксплота. (1 б.) *Дополнительно: Добавьте название графика и подпись* (0.5 б.)

Разное

1. Постройте “леденцовый график” (lollipop-plot) для количества набранных студентами 5-го курса баллов за весь учебный год (по оси ординат – id студента, по оси абсцисс – итоговый балл). Отсортируйте студентов в порядке убывания итогового балла. Раскрасьте точки на “леденцах” в зависимости от сердцевины волшебной палочки. Палочки с сердечной жилой дракона должны быть красного цвета, с пером феникса – желтого, с волосом единорога – серого. (1 б.)
2. Постройте гистограмму распредления баллов за экзамен по астрономии. Выделите цветом факультет Слизерин. Примените 18-й кегль к тексту на осях x, y и легенды. Название оси y и легенды запишите 20-м кеглем, оси x – 22-м. Измените название оси y на “Number of students”. (1 б.)
3. На лекции мы использовали комбинацию theme_bw(), и созданной нами theme_custom, чтобы одновременно сделать фон белым и увеличить шрифт. Модифицируйте theme_custom таким образом, чтобы она и выполняла свои прежние функции, и делала фон белым без помощи theme_bw(). Примените новую кастомную тему к графику, полученному в последнем пункте блока по боксплотам (1.5 б).

Фасетирование

Существует еще один способ визуализировать несколько распределений на одном графике – фасетирование (семейство функций `facet_`). При создании фасетов вы получаете на одном графике несколько субграфиков, например, вот так:

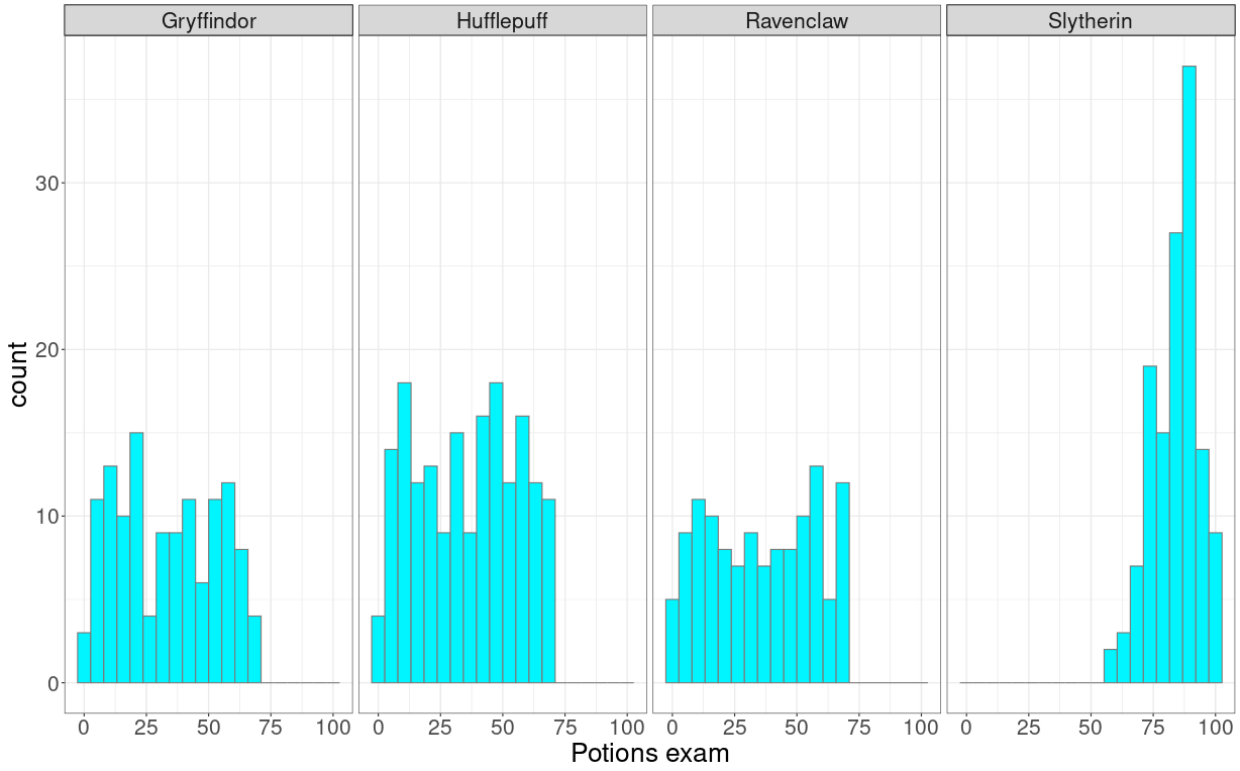


Фасетирование по строкам

Два основных подхода к фасетированию – функции `facet_grid()` и `facet_wrap()`.
`facet_grid()` в простейшем случае принимает формулу вида `x~y`, где x показывает по какой переменной мы фасетируем по строкам, а y – по какой переменной мы фасетируем по столбцам. Если мы хотим сделать фасет только по одному измерению, на месте второго мы ставим точку.

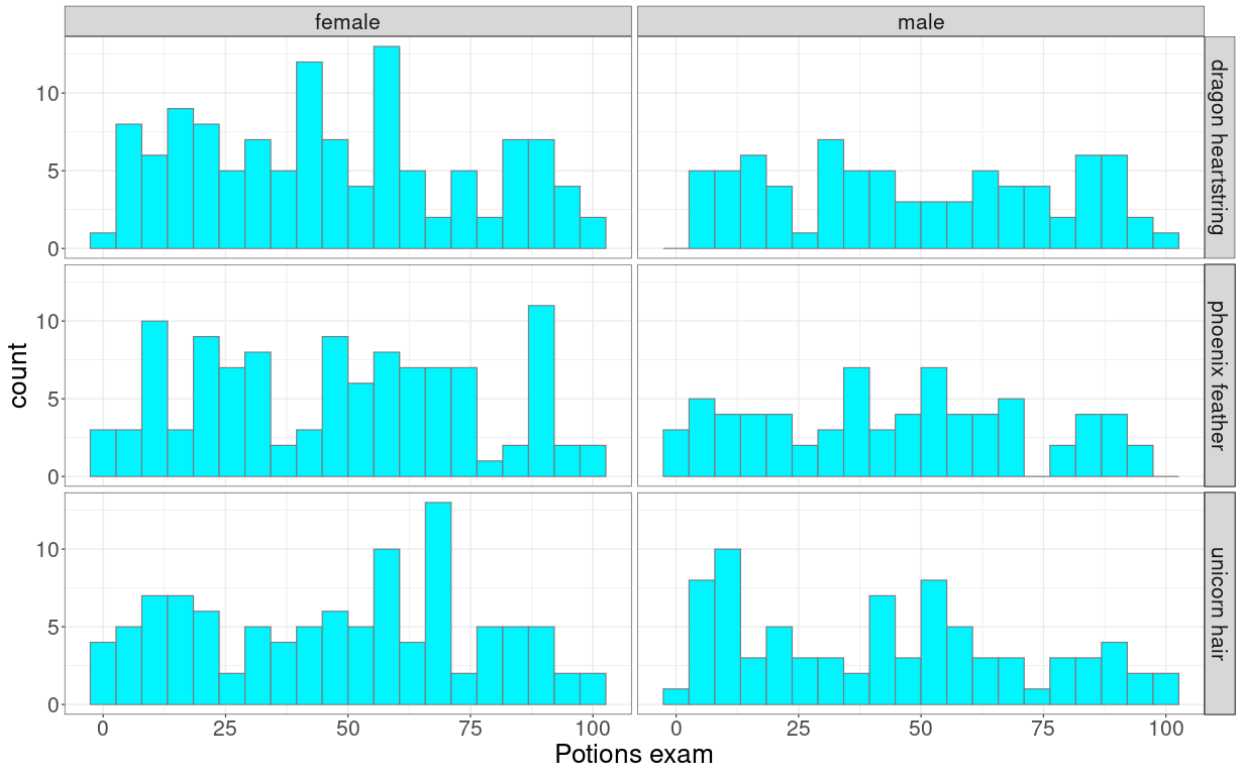
Например, иллюстрация выше была создана при помощи кода `facet_grid(house~.)` (разбили графики на “строки” по переменной house, по “столбцам” не разбивали).

Строчка кода `facet_grid(.~house)` создала бы следующий график:



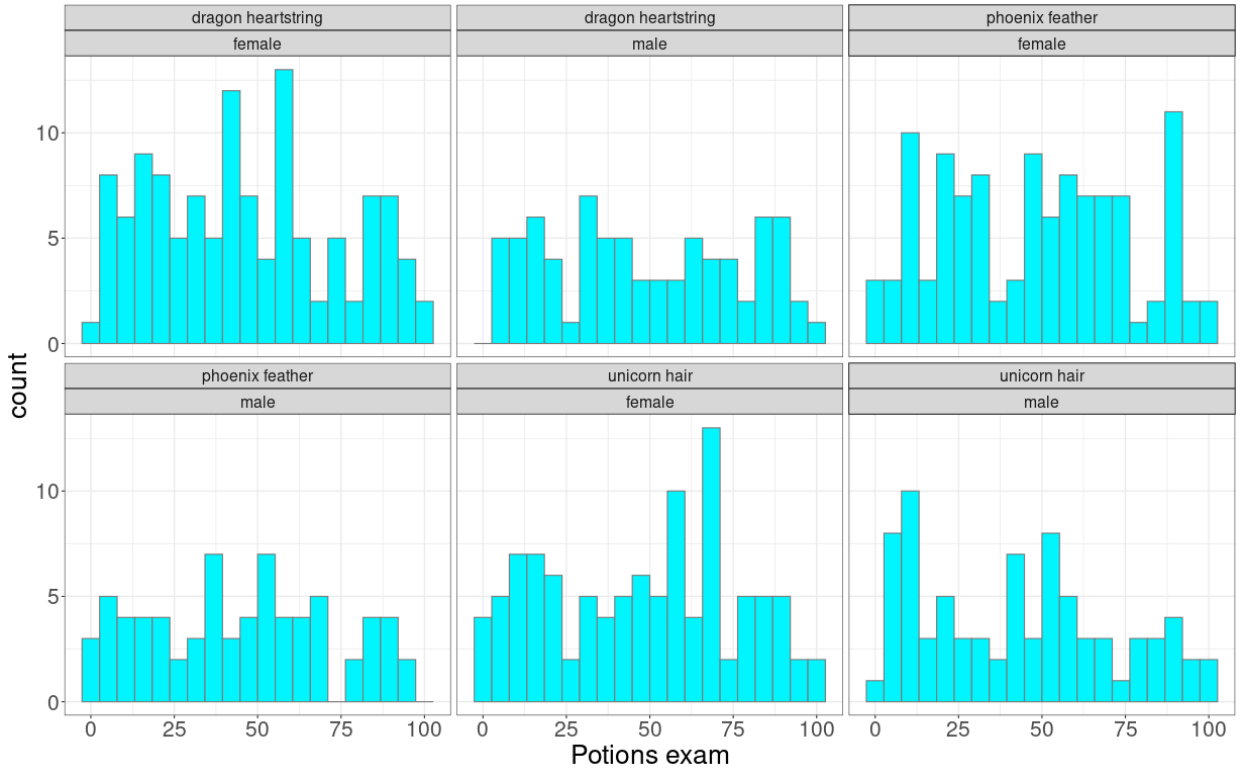
Фасетирование по столбцам

Если бы мы хотели добавить еще одну группирующую переменную (например, пол) в фасет и выполнили код `facet_grid(wandCore~sex)`, результат бы выглядел следующим образом.



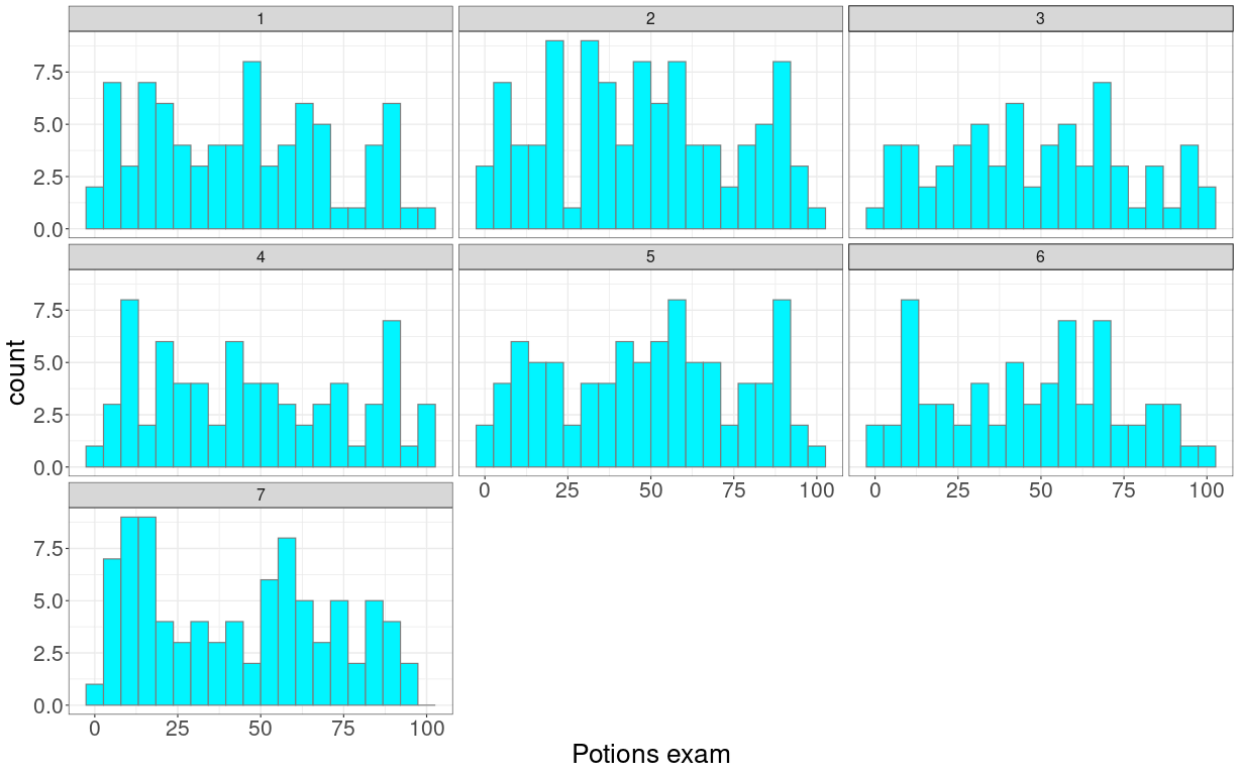
Фасетирование по строкам и столбцам

Второй способ разбить исходное изображение на субграфики – функция `facet_wrap()`. Она пытается оптимизировать занимаемое пространство и максимально “оквадратить” итоговый график. Синтаксис следующий: `facet_wrap(vars(x, y, z))`, где x, y, z – 1 и более группирующая переменная. Например, фасетирование по тем же переменным, что и в прошлом примере с использованием facet_wrap(vars(wandCore, sex)) будет выглядеть так:



facet_wrap

а `facet_wrap(vars(course))` так:



1. Напишите, какой, по вашему мнению, способ фасетирования (по строкам или по столбцам) лучше использовать для визуализации гистограммы. Почему? А какой для визуализации violin-plot? Почему? Можно ли вывести общее правило? (1.5 б)
2. Постройте гистограмму для результата любого выбранного вами экзамена, кроме зельеварения. Настройте оптимальное на ваш взгляд число столбцов гистограммы. Выполните фасетирование по курсу. Постарайтесь, чтобы график был по возможности компактным. (1 б.).
3. Отобразите на одном графике распределение плотности вероятности для оценки студентов на экзамене по защите от темных искусств и на экзамене по травологии. Раскрасьте их в любые выбранные вами цвета, постарайтесь, чтобы оба распределения отображались целиком. Примените тему из 3-го пункта блока “Разное”. Сделайте фасетирование по полу (1 б.).

Ссылка на документацию ggplot2.