BI\_retraining\_basic\_datavis\_2024

B.V. Sotnikov

2024-09-11

# Чтение данных

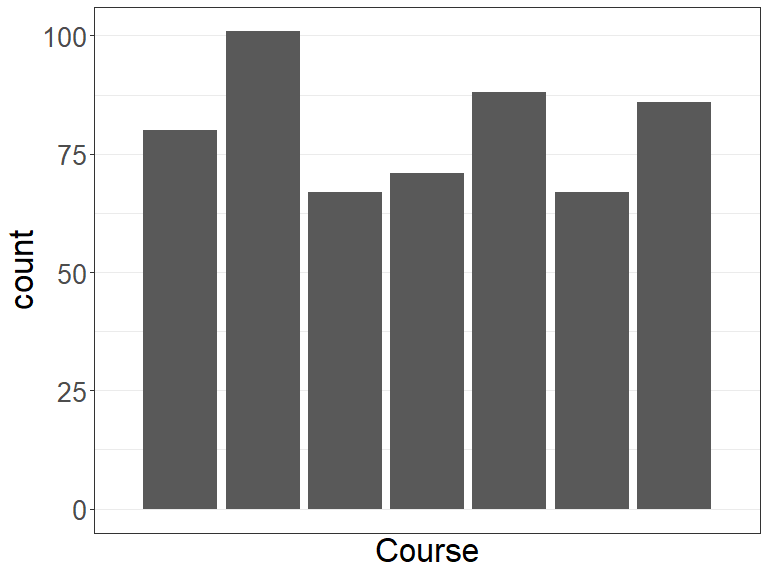
hogwarts <- read\_csv("../data/raw/Visualization\_Data/hogwarts\_2024.csv")  
hogwarts |> head()

## # A tibble: 6 × 60  
## id house course sex wandCore bloodStatus result Defence against the …¹  
## <dbl> <chr> <dbl> <chr> <chr> <chr> <dbl> <dbl>  
## 1 1 Ravencl… 4 fema… unicorn… half-blood 94 73  
## 2 2 Hufflep… 5 male phoenix… half-blood 33 38  
## 3 3 Ravencl… 4 fema… dragon … half-blood 137 52  
## 4 4 Hufflep… 2 male phoenix… half-blood 27 50  
## 5 5 Hufflep… 2 fema… phoenix… half-blood 67 47  
## 6 6 Gryffin… 6 male phoenix… muggle-born 126 44  
## # ℹ abbreviated name: ¹​`Defence against the dark arts exam`  
## # ℹ 52 more variables: `Flying exam` <dbl>, `Astronomy exam` <dbl>,  
## # `Herbology exam` <dbl>, `Divinations exam` <dbl>, `Charms exam` <dbl>,  
## # `History of magic exam` <dbl>, `Arithmancy exam` <dbl>,  
## # `Care of magical creatures exam` <dbl>, `Muggle studies exam` <dbl>,  
## # `Study of ancient runes exam` <dbl>, `Transfiguration exam` <dbl>,  
## # `Potions exam` <dbl>, week\_1 <dbl>, week\_2 <dbl>, week\_3 <dbl>, …

# Графики 1: Столбчатые диаграммы

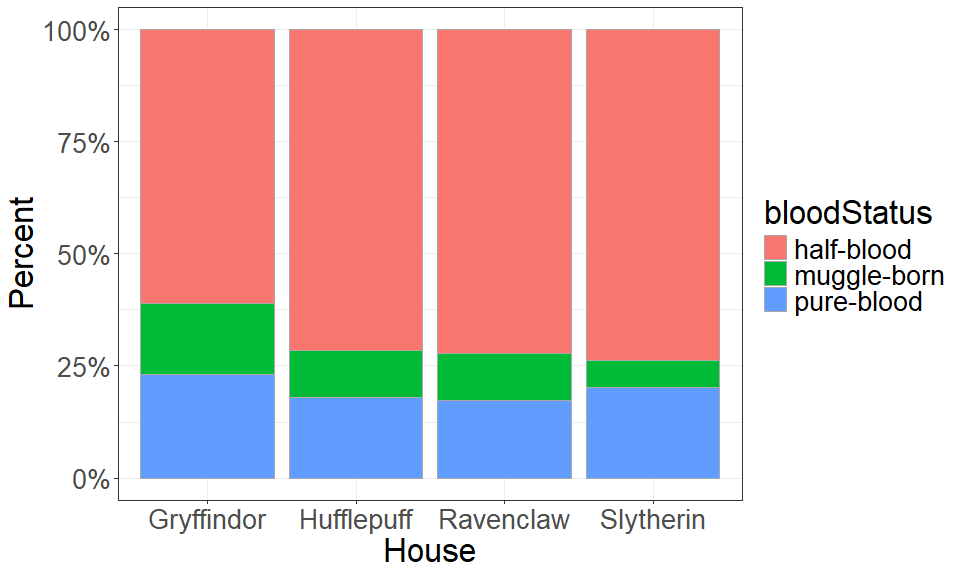
### Задание 1

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).



### Задание 2

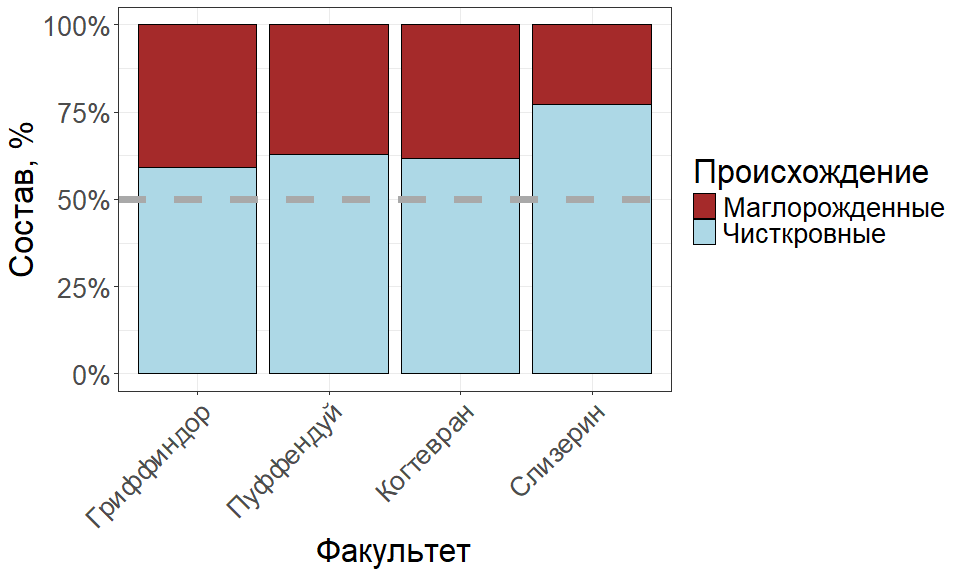
Создайте новый барплот, отражающий распределение числа студентов по факультету. Добавьте на график вторую факторную переменную – происхождение (bloodStatus). Модифицируйте при помощи аргумента position графика так, чтобы каждый столбец показывал распределение факультета по чистоте крови в долях. Примените произвольную тему. Запишите текстом в rmd-документе, какой вывод можно сделать из графика? (1 б).



На графике наблюдаем тренд, что “полукровок” большинство среди учащихся студентов Хогвартса. Первая интерпретация, которая приходит в голову - это то, что маги склонны выбирать себе супругов не магов :)

### Задание 3

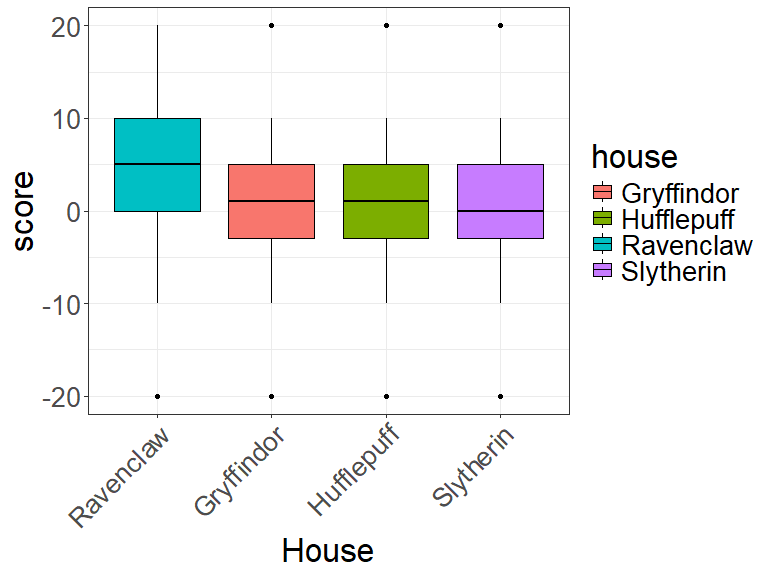
Модифицируйте датасет таким образом, чтобы в нем остались только чистокровные (pure-blood) и маглорожденные студенты (muggle-born). Создайте на основе этих данных график из пункта 2. Добавьте горизонтальную пунктирную линию произвольного цвета на уровне 50%. Дайте осям название на русском языке (1б). Дополнительно: переименуйте на русский язык категории легенды pure-blood и muggle-born (0.5 б).



# Графики 2: Бок-плоты

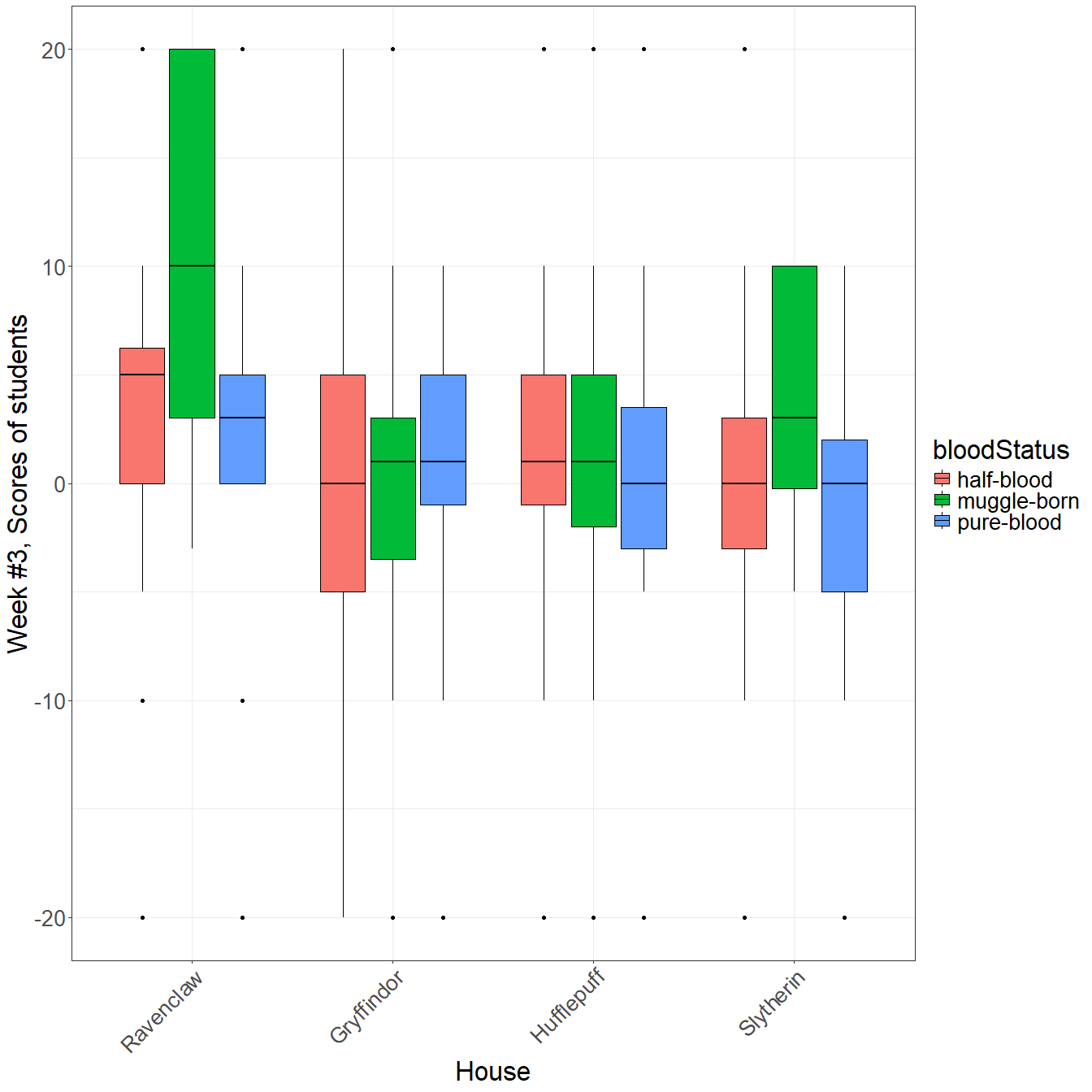
## Задание 1

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.)



## Задание 2

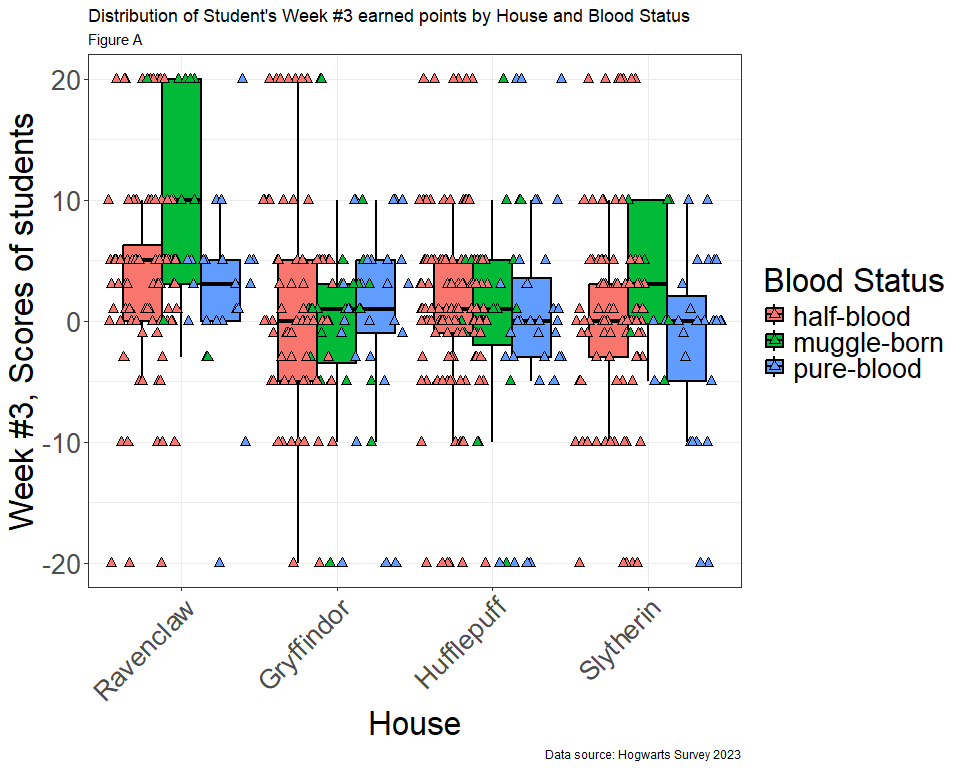
Добавьте отображение разными цветами для происхождения студентов (bloodStatus). Добавьте на боксплот вырезку (notch). Настройте для данного чанка размер изображения 14:14 дюймов. Приведите названия осей к корректному виду. (1 б.)



Все-таки я решил сделать версию без ‘notch=TRUE’. С нотчом график не корректно отображается, т.к. в группе “muggle-born”, слишком мало наблюдений и КИ для медианы шире, чем IQR, из-за чего возникает ошибка.

## Задание 3

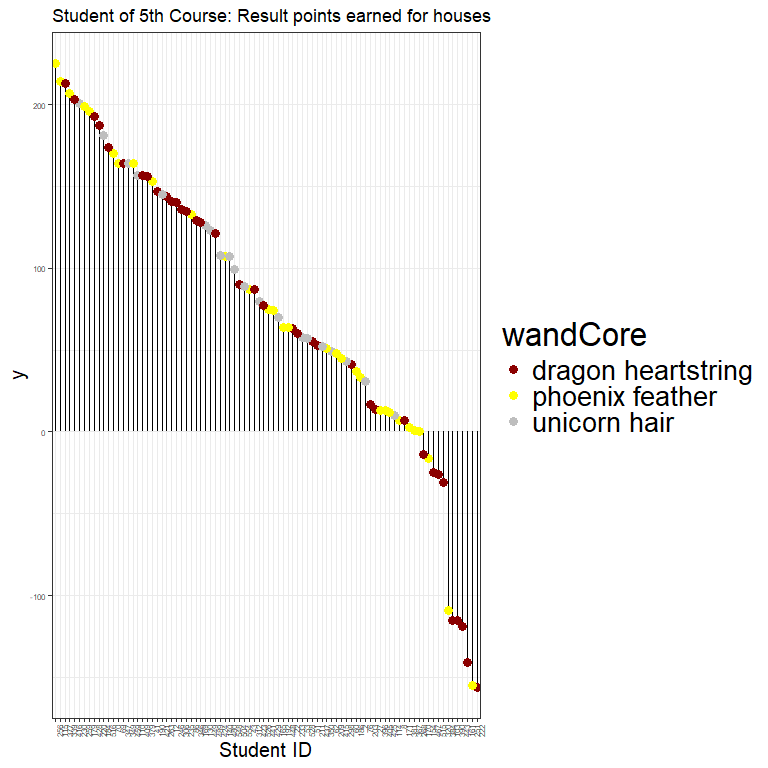
Добавьте на график джиттер-плот. Удалите отображение выбросов у боксплота. Видоизмените по своему вкусу толщину линий и ширину боксплота. (1 б.) Дополнительно: Добавьте название графика и подпись (0.5 б.).



# Графики 3: Разное

## Задание 1

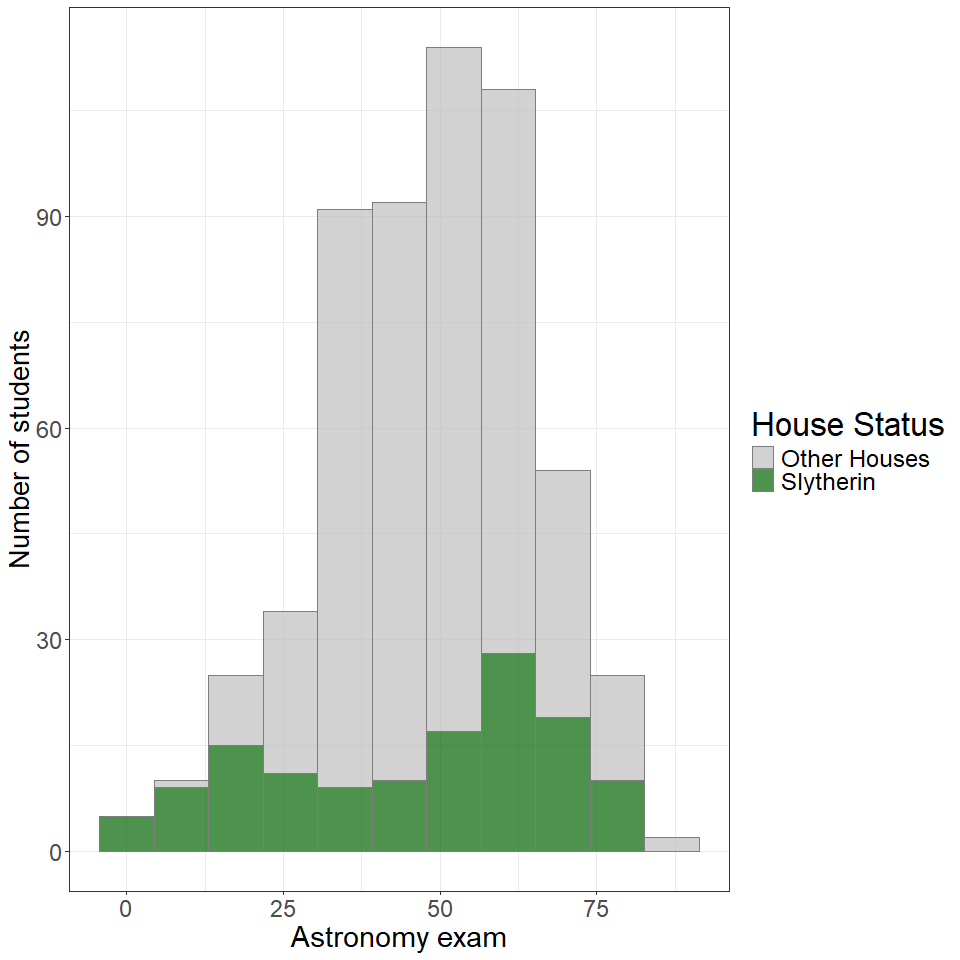
Постройте “леденцовый график” (lollipop-plot) для количества набранных студентами 5-го курса баллов за весь учебный год (по оси ординат – id студента, по оси абсцисс – итоговый балл). Отсортируйте студентов в порядке убывания итогового балла. Раскрасьте точки на “леденцах” в зависимости от сердцевины волшебной палочки. Палочки с сердечной жилой дракона должны быть красного цвета, с пером феникса – желтого, с волосом единорога – серого. (1 б.)



Что-то айдишнике съехали немного, но совсем понять не смог, почему так.

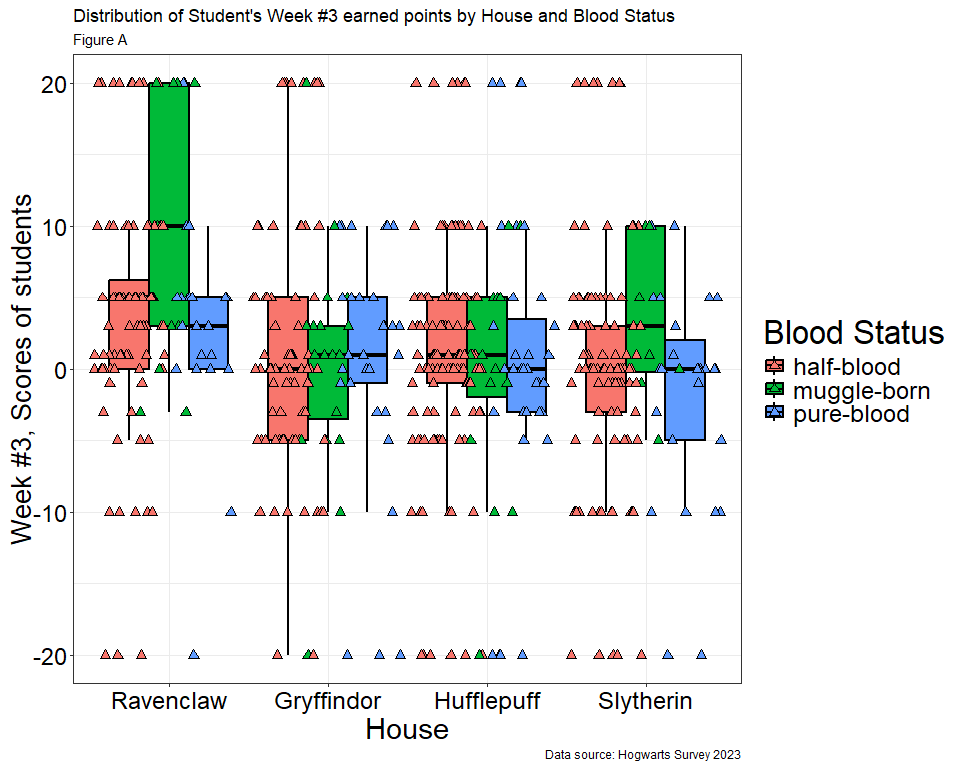
## Задание 2

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.)



## Задание 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).



# Фасетирован

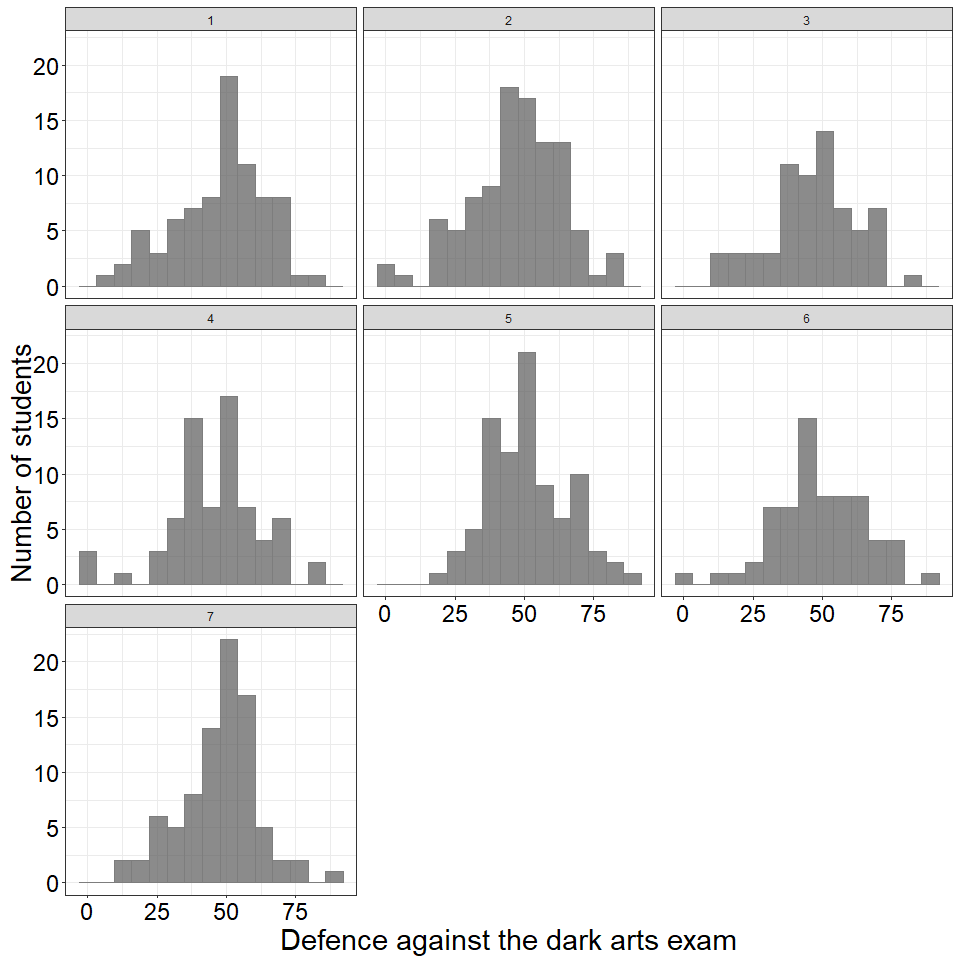
## Задание 1

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)

#### Ответ: Считаю, что для гистограмм фасетирование по строкам даёт более удобную репрезентацию данных, т.к. у нас х-ось общая для графиков, поэтому легче визуально оценить различия между распределениями какого-то показателя между группа. Наоборот, для виолин-плотов удобнее фасетирование по столбцам, причина та же - общая у-ось, позволяющая легче увидеть различия или их отсутствия между группами. Т.о. можем вывести общее правило для графиков: фасетирование лучше делать так, чтобы “основная” ось (ось, показывающая распределения какого-либо фактора) была общая между графиками, построенными с помощью такого метода. Однако, стоит заметить, что если кол-во групп больше 4-5, то фасетирование только по строкам/столбцам может быть не очень удачным решением, т.к. график получится очень большим/узким, что снизит читаемость.

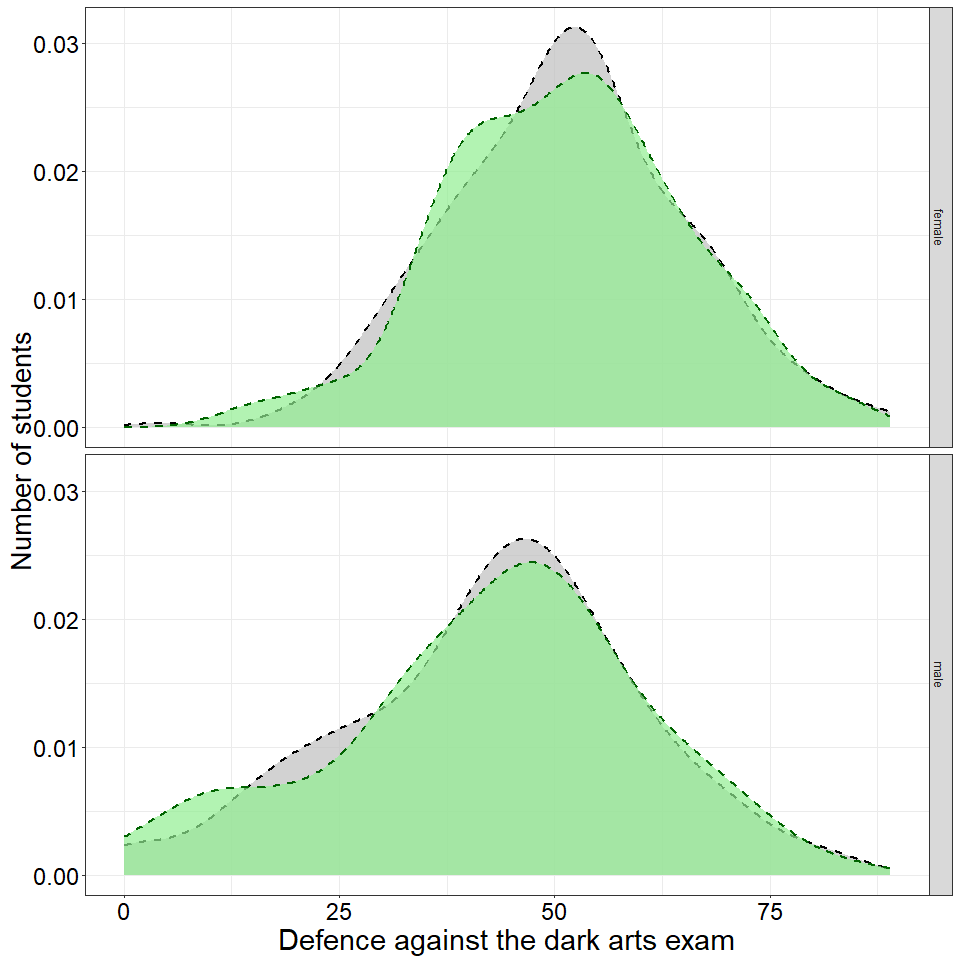
## Задание 2

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.).

 Здесь использовал facet\_wrap(vars(course)), чтобы график получился компактным

## Задание 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.).

 Готово!