

P2 - Alex Carriero

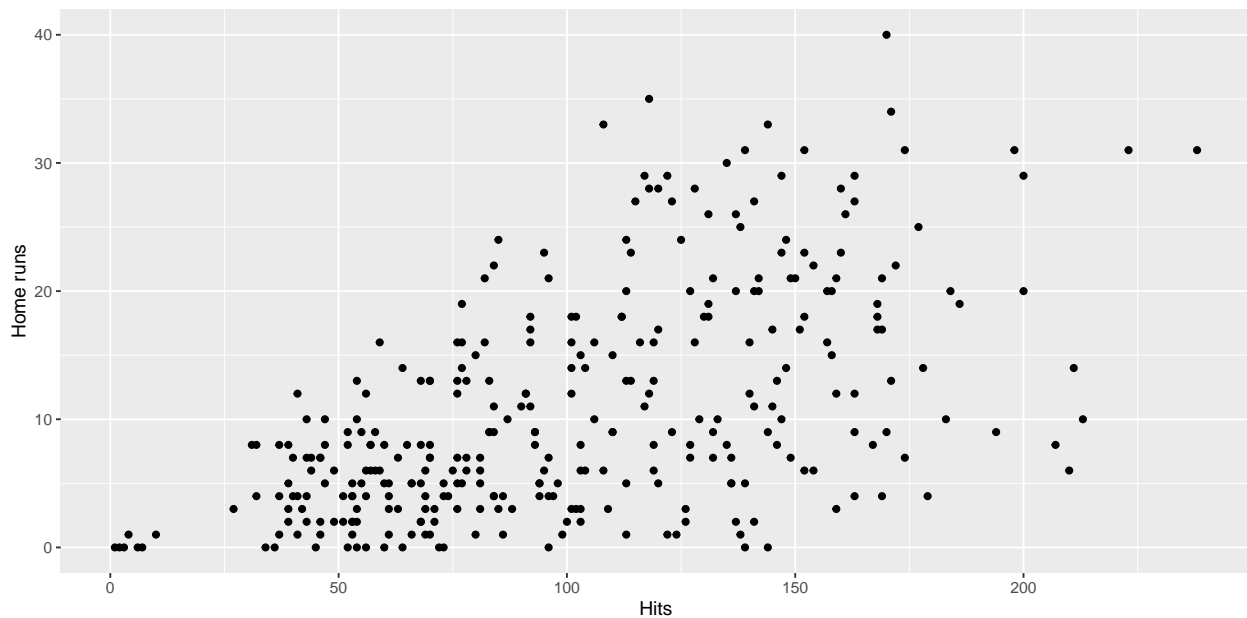
Alex Carriero

19/09/2022

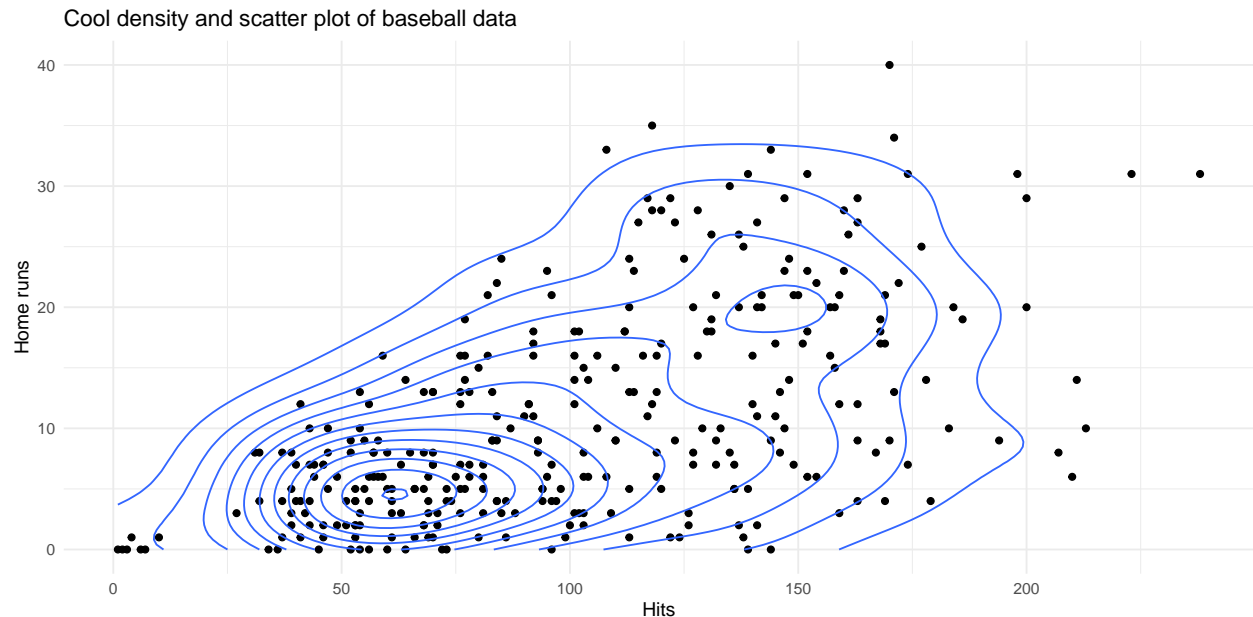
```
# Libraries
library(ISLR)
library(tidyverse)
```

```
# Intro to GGplot using Hitters dataset
homeruns_plot <-
  ggplot(Hitters, aes(x = Hits, y = HmRun)) +
  geom_point() +
  labs(x = "Hits", y = "Home runs")
```

```
homeruns_plot
```



```
# layered plots
homeruns_plot +
  geom_density_2d() +
  labs(title = "Cool density and scatter plot of baseball data") +
  theme_minimal()
```



1. Name the aesthetics, geoms, scales, and facets of the above visualisation. Also name any statistical transformations or special coordinate systems.

- Aesthetics: $x = \text{Hits}$, $y = \text{Home Runs}$
- Geoms:
 - `geom_point()` used to create scatter plot
 - `geom_density_2d()` used to perform a 2D kernel density estimation and display the results with contours.
- Scales: none
- Facets: none

2. Run the code below to generate data. There will be three vectors in your environment. Put them in a data frame for entering it in a `ggplot()` call using either the `data.frame()` or the `tibble()` function. Give informative names and make sure the types are correct (use the `as.()` functions). Name the result `gg_students`.

```
# Generate data
set.seed(1234)
student_grade <- rnorm(32, 7)
student_number <- round(runif(32) * 2e6 + 5e6)
programme <- sample(c("Science", "Social Science"), 32, replace = TRUE)

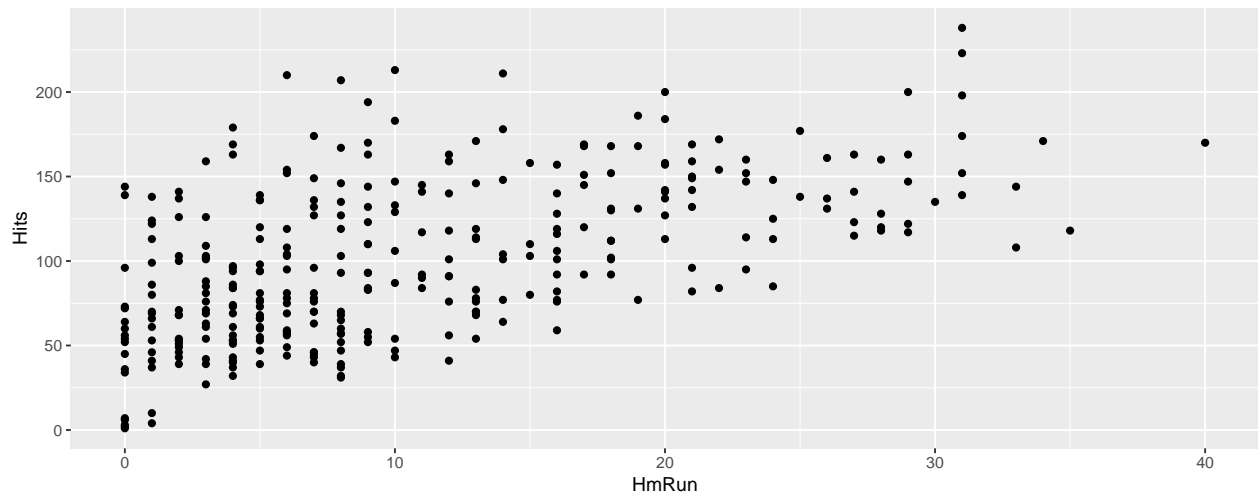
# Store in data frame

gg_students <- as.data.frame(cbind(student_number, student_grade, programme))
gg_students <- gg_students %>%
  mutate(student_number = as.numeric(student_number),
         student_grade = as.numeric(student_grade),
         programme = as.factor(programme))

# gg_students
```

3. Plot the first homeruns_plot again, but map the Hits to the y-axis and the HmRun to the x-axis instead.

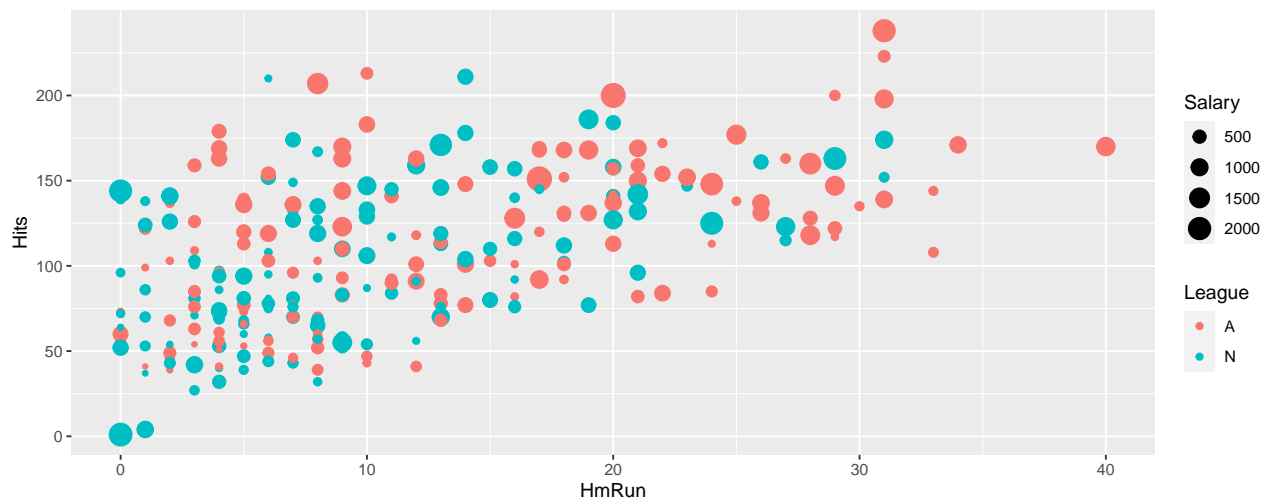
```
ggplot(Hitters, aes(x= HmRun, y = Hits)) +  
  geom_point()
```



4. Recreate the same plot once more, but now also map the variable League to the colour aesthetic and the variable Salary to the size aesthetic.

```
summary(Hitters$League)  
summary(Hitters$Salary)  # there are 59 missing salary values  
  
Hitters <- Hitters%>%  
  na.omit(Hitters$Salary)
```

```
new_plot <- ggplot(Hitters, aes(x= HmRun, y = Hits)) +  
  geom_point(aes(color = League, size = Salary))  
  
new_plot
```



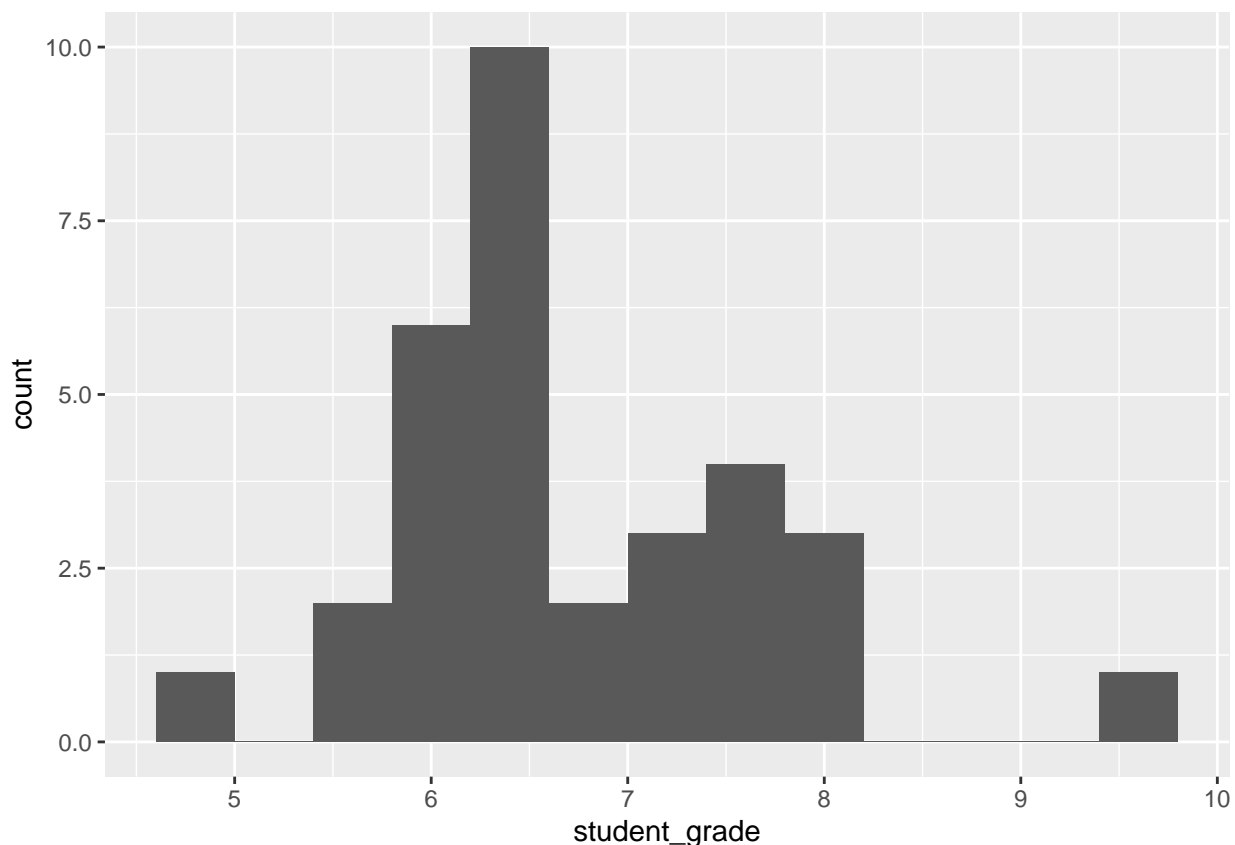
5. Look at the many different geoms on the reference website.

There are two types of geoms:

- geoms which perform a transformation of the data beforehand, such as `geom_density_2d()` which calculates contour lines from x and y positions.
- geoms which do not transform data beforehand, but use the aesthetic mapping directly, such as `geom_point()`.

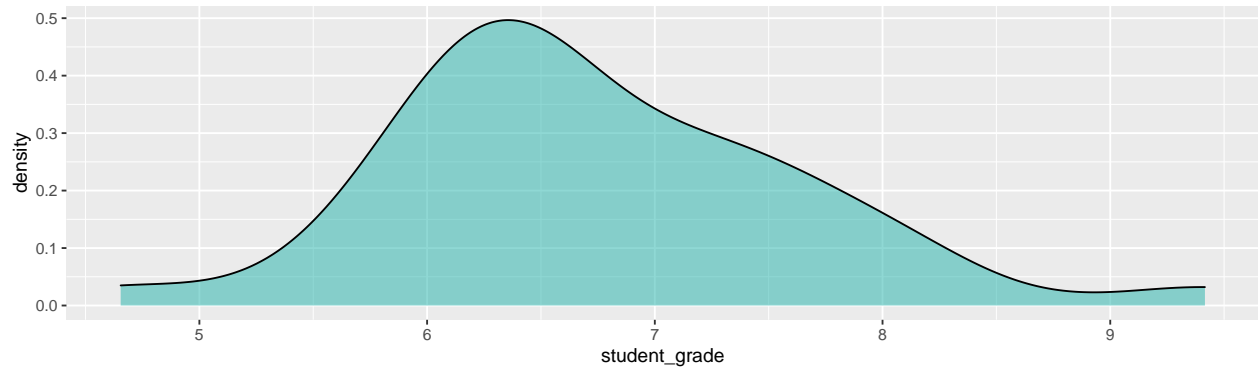
6. Use `geom_histogram()` to create a histogram of the grades of the students in the `gg_students` dataset. Play around with the `binwidth` argument of the `geom_histogram()` function.

```
gg_students %>%  
  ggplot(aes(x= student_grade))+  
  geom_histogram(binwidth=0.4)
```



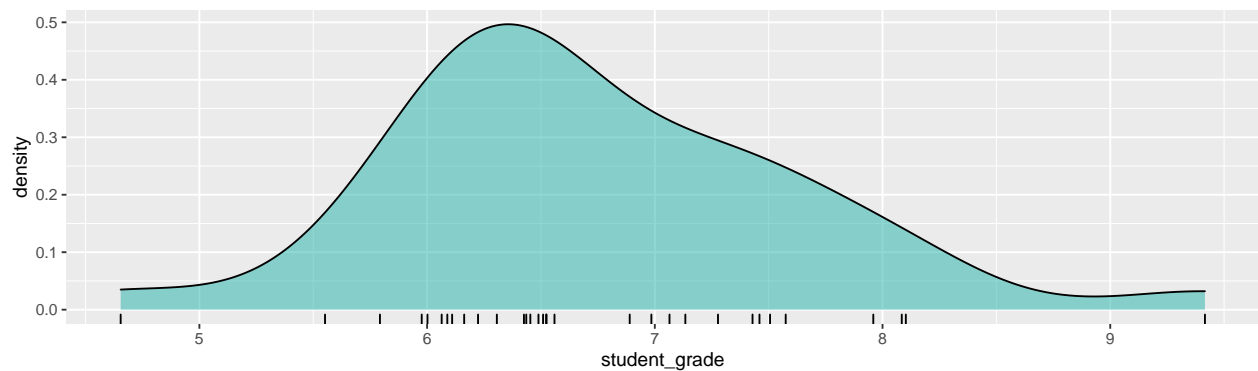
7. Use `geom_density()` to create a density plot of the grades of the students in the `gg_students` dataset. Add the argument `fill = "light seagreen"` to `geom_density()`.

```
d <- gg_students %>%  
  ggplot(aes(x= student_grade))+  
  geom_density(fill = "light seagreen", alpha= 0.5)  
d
```



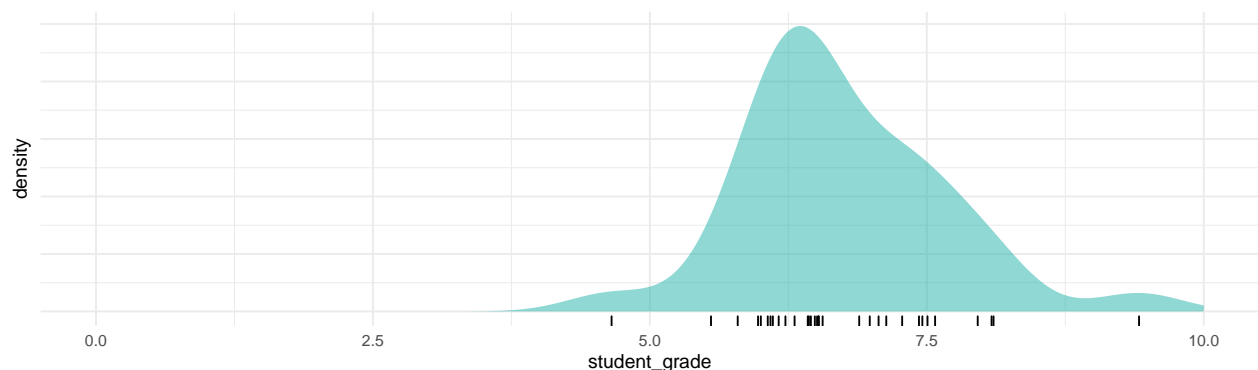
8. Add rug marks to the density plot through `geom_rug()`. You can edit the colour and size of the rug marks using those arguments within the `geom_rug()` function.

```
d +  
  geom_rug(color = "black", size = 0.5)
```



9. Increase the data to ink ratio by removing the y axis label, setting the theme to `theme_minimal()`, and removing the border of the density polygon. Also set the limits of the x-axis to go from 0 to 10 using the `xlim()` function, because those are the plausible values for a student grade.

```
gg_students %>%  
  ggplot(aes(x= student_grade))+  
  geom_density(fill = "light seagreen", alpha= 0.5, color = NA)+  
  geom_rug(color = "black", size = 0.5)+  
  theme_minimal()+  
  theme(axis.text.y=element_blank()) +  
  xlim(0,10)
```



10. Create a boxplot of student grades per programme in the `gg_students` dataset you made earlier: map the programme variable to the x position and the grade to the y position. For extra visual aid, you can additionally map the programme variable to the fill aesthetic.
11. What do each of the horizontal lines in the boxplot mean? What do the vertical lines (whiskers) mean?
12. Comparison of distributions across categories can also be done by adding a fill aesthetic to the density plot you made earlier. Try this out. To take care of the overlap, you might want to add some transparency in the `geom_density()` function using the `alpha` argument.
13. Create a bar plot of the variable Years from the Hitters dataset.