Exercises - Visualizing Distributions

Data Visualization

Instructions and tips:

- Open the template file for this exercise class and work in that file. The template contains some starting code and all the questions text.
- Use the R help files to see the options that are available for a function. You can run `?functionname` (e.g. `?mean`) in your console, or use the bottom right screen in RStudio to look up help files.
- To save some time you can copy parts of code from your answers to a first subquestion, and paste and adjust it in the next subquestion as many are sequential.

Histograms

Exercise 1: Create bad histograms

In this exercise we will use the data on cow butterfat (`cows.rda`) that was used in the lecture.

a. Make a histogram of cow butterfat.

b. Purposefully obscure interesting or important information by playing with bin width and/or number.

c. Make it **ugly**.

Exercise 2: Bin settings

Use the `diamonds` data set that is available in `R` to perform this exercise.

Create a histogram which displays the distribution of price and modify several bin settings in `geom_histogram()` such as the number of bins, binwidth, and the use of arguments `center` and `boundary`. What is the result of each modification?

Exercise 3: Multipanel histograms

Continue working with the `diamonds` data set that is available in `R`. Create multipanel histograms that display the price distribution based on diamond cut.

a. A histogram with axes which both scales (x and y axis) are fixed. What insight can we get from fixing the scales?

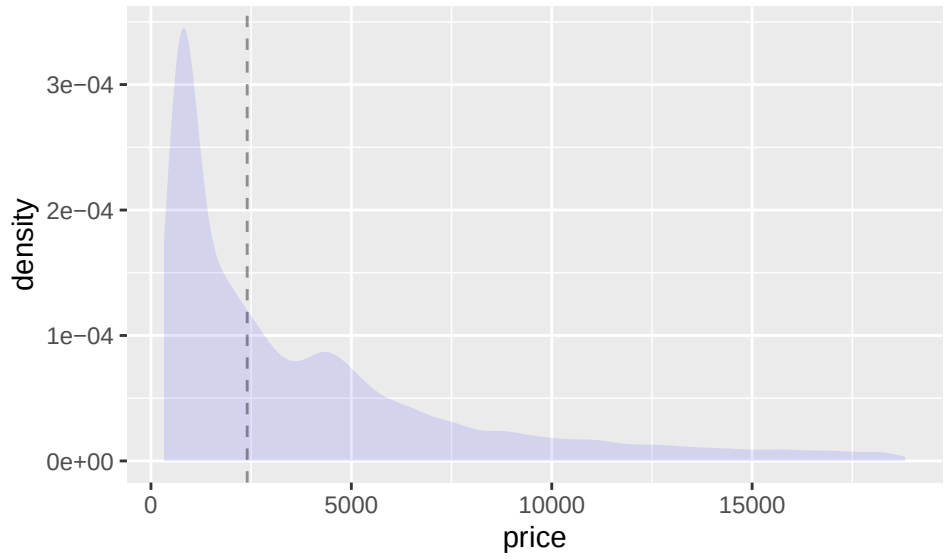
b. A histogram in which only the scale of the x-axis is fixed. What information do we gain/loose by only fixing the scale of the x-axis for each diamond cut?

c. A histogram with different bin colors based on diamond cut. Would giving colors to the bins give more insight to the data? Explain your answer.

Density functions

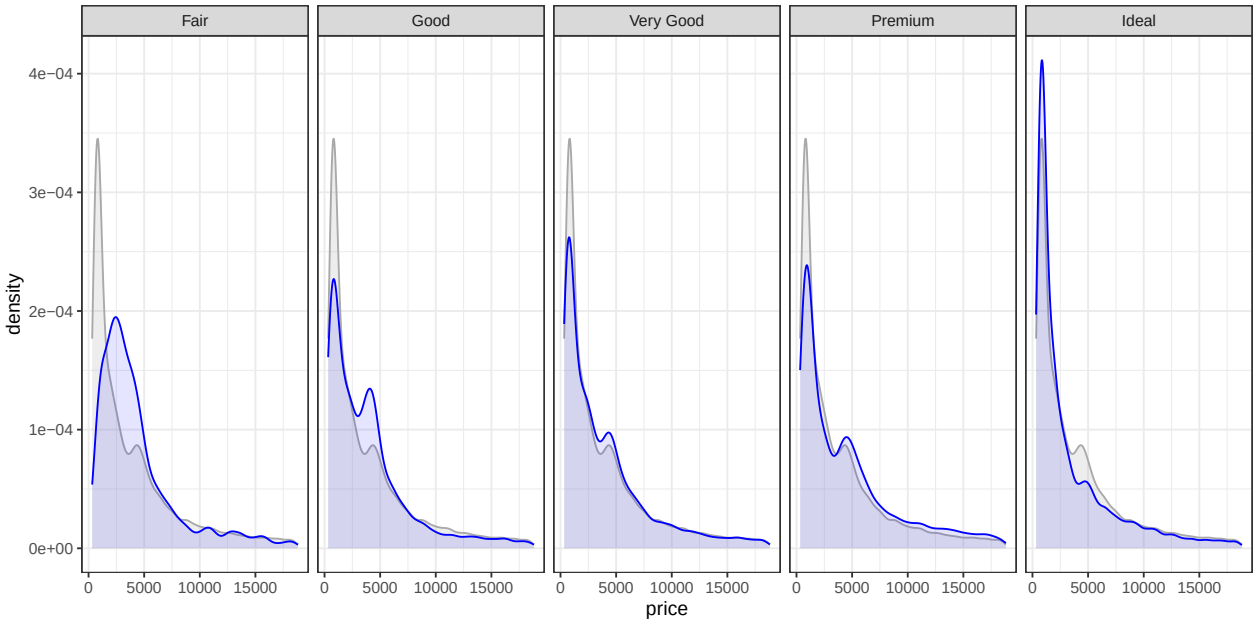
Exercise 4

Again, continue working with the `diamonds` data set that is available in `R`. Create the density plot for price of diamonds that is shown below, i.e. add a line which indicates the median, color the area of the density in blue, and make the area slightly transparent.



Exercise 5

Again, continue working with the `diamonds` data set that is available in `R`. Create a multipanel and multilayered density plot which compares the distribution of price of each diamond cut to the overall price distribution. Like the one below.

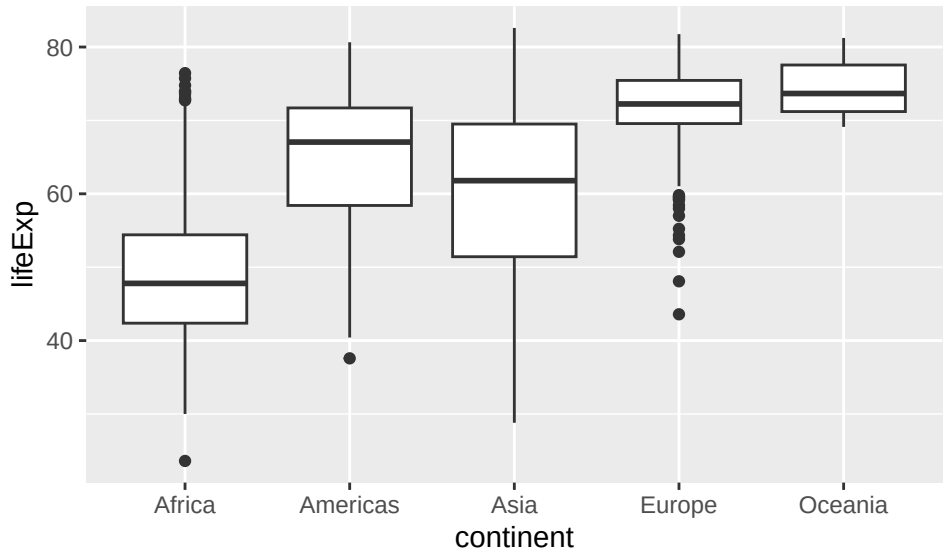


Boxplots

Exercise 6

Use the `gapminder` data from the `gapminder` package for this exercise.

a. Create boxplots that display the distribution of life expectancy at birth for each continent, as in the figure below.



a. Flip the axes of the above exercise and color the boxes light blue.

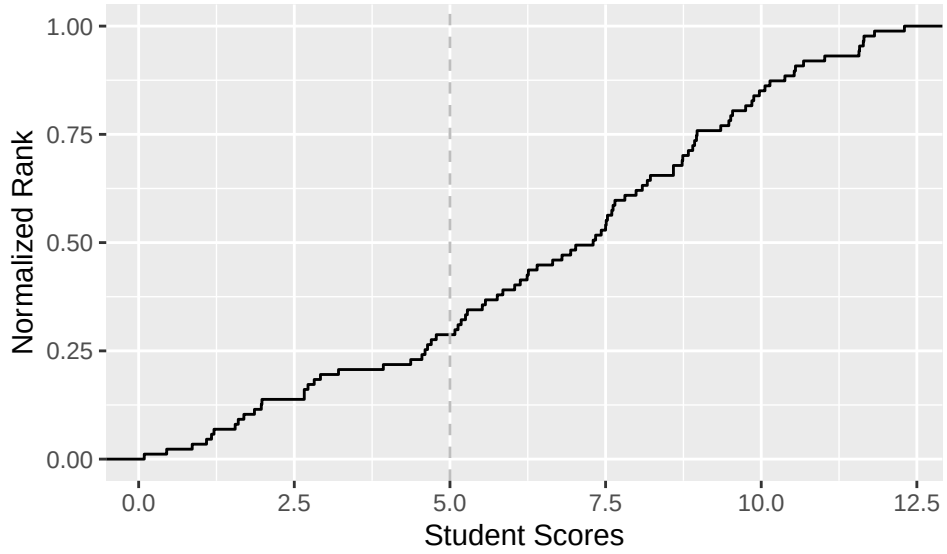
b. Fill the boxes with different colors and suppress the color legend.

c. Include the individual data points as jitter on top of the bloxplots. Remove the outliers so that they are not mixed up with the data points. Try various transparencies so that both the boxplot and the data points are visible.

Empirical Distribution Plot

Exercise 7

Using the `end-of-year.csv` file and `stat_ecdf()` make a plot showing the proportion of students with scores below 5.

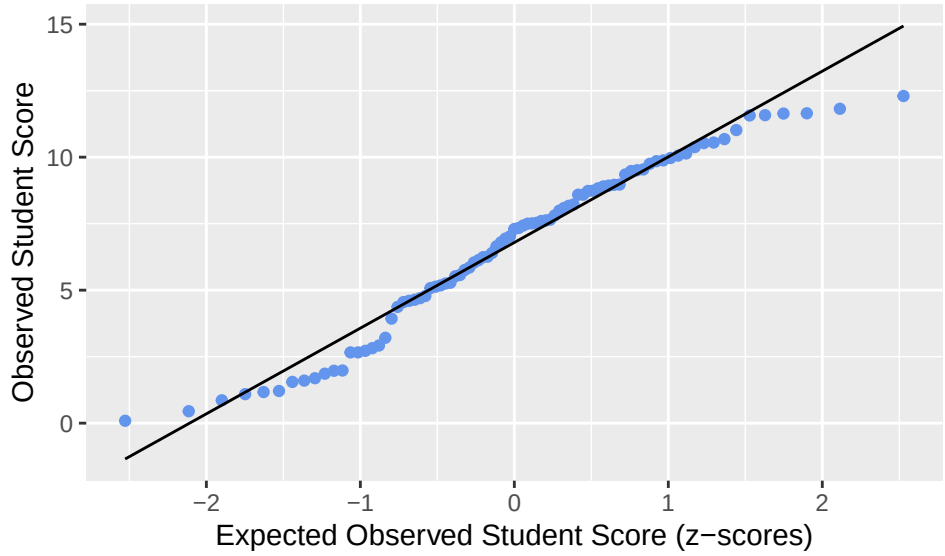


Quantile-Quantile (Q-Q) Plots

Exercise 8

Using the student score data (`end-of-year.csv`) determine to what extent the observed score data follow Gaussian distribution.

Hint: Use `stat_qq()` and `stat_qq_line()`



Challenges

Meteorological data from Central Park, NYC

The dataset `central-park.csv` contains daily meteorological data from Central Park, NYC, in 2022. Pick variables that appeal to you and display their distribution throughout the year in a logical and meaningful way.

Remake the plot

Using the student score data (`end-of-year.csv`), remake the plot below which shows the the distribution of end of year assessment scores.