Visualization, Homework

Data Science Team

Due Tuesday July 5 at 10am

Please upload the `PDF` that you obtain by knitting the `Rmd` file that contains your `R` code and your text answering other questions. So this uploaded file will also show any output that `R` produces in addition to your code.

1. Diamonds data

Note: though the data for this question are found in the library `ggplot2`, you may use either base R (built-in) graphical features or make ggplots.

(a) Load in the data with command

```
data(diamonds, package = "ggplot2")
```

(b) Is there a relationship between the weight of the diamond (as measured by `carat`) and `price`? Draw a plot and comment.

(c) Is there a relationship between `carat` and `price` for diamonds for the subset $1 \leq carat \leq 2$? To answer this question, you can select those diamonds with the `subset` function using the boolean vector `carat >=1 & carat <=2`. Look up what the boolean operator `&` does so you understand this way of extracting a subset of the data, which is common and useful in R.

(d) Make a barplot of `clarity` and interpret it.

(e) Fill each bar in (d) with different colors corresponding to `cut`.

(f) Facet the plot in (c) by `color`. What does this show about the relationship among the variables?

2. A data set that piques your curiosity

Browsing the web, identify one dataset that piques your curiosity (feel free to start from one of the sites mentioned in lecture).

(a) Specify which dataset you are going to be working with, providing a link to the original source.

(b) Load the data in `R`.

(c) Describe the variables involved and present summary statistics and graphical displays.

(d) Formulate a question that you believe the data can help you answer. You do not need to answer the question.

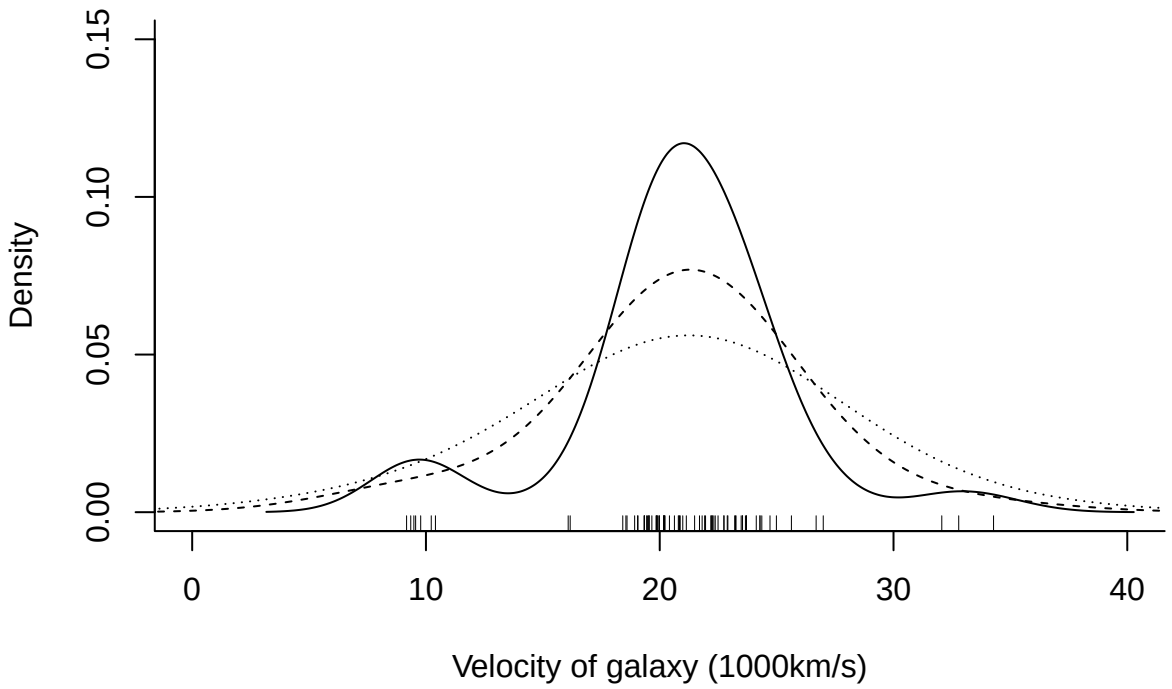
3. How many clusters of galaxies?

This exercise uses data analyzed in Roeder (1990) Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies available in the `R` library `MASS`. The data are the measured velocities in km/second of 82 galaxies from the Corona Borealis region. You may need to install the package "MASS", e.g. via "install.packages()".

```
library(MASS)
gal <- galaxies/1000
plot(x = c(0, 40), y = c(0, 0.15), type = "n", bty = "l",
     xlab = "Velocity of galaxy (1000km/s)", ylab = "Density")

rug(gal)                                    # add a 'rug' (ticks along x axis)

lines(density(gal, bw = 6), lty = 3)        # bw means bandwidth
lines(density(gal, bw = 4), lty = 2)
lines(density(gal, bw = 2), lty = 1)        # lty controls y's appearance
```



The number of modes (i.e. local maxima of the density) in this data is of scientific interest. If you're curious, take a look at the referenced article to understand, but here is a short summary: A theory of how the universe formed predicts the existence of clusters of galaxies. If galaxies travel at similar speeds, that means that the galaxies are clumped. This suggests that multimodal data corresponds to multiple clusters of galaxies. A unimodal distribution, by contrast, is what one would expect if there were no clusters and the data were just an artifact of how the galaxies were sampled.

The plot above includes three different density estimates for the data, using three different levels of smoothing, specified by the parameter `bw` (an abbreviation for bandwidth). The bandwidth controls how much the data are 'smoothed'.

(a) Which bandwidth provides the clearest evidence that the galaxies are clustered?

(b) Adapt the code above so that you add a fourth line that is less 'smooth' than any of the ones there.

(c) Imagine you are writing an article for a popular scientific magazine and the editor says that you may only include one density line. Which bandwidth would you choose to best represent the data and why?

4. Will my flight be on time?

Load the package `nycflights13`. This package contains data on flights that departed from the three major NYC airports (LGA, EWR, and JFK) in 2013. Start by loading the main data and merging it with airline carrier data like so:

```
library(pacman)
p_load(nycflights13, dplyr)
flights <- inner_join(flights, airlines, by="carrier")
flights <- flights[,-20]                                # drop extra copy of 'name'
colnames(flights)[19] <- "name"                         # simplify variable name
```

You may need to install the package "pacman", e.g. via "install.packages()".

(a) After getting an overview of the data with `glimpse`, create a variable that captures the delay of each flight by adding the departure delay to the arrival delay. Plot its density.

(b) Which carrier has the best record around the holidays? Create a data visualization that shows delays by carrier between December 15th and January 5th. Make sure your visualization conveys how well the carrier does on average as well as a sense of uncertainty of the estimate, for example by creating boxplots. (Similarly to above, you can extract the relevant flights using `(month == 12 & day >= 15) | (month == 1 & day <=5)`. Look up the logical operator `|` and understand what it does.)

(c) Create a plot that shows whether and how the distance between airports affects the delay.

(d) Briefly describe a hypothesis you have about which flights are most likely to be delayed (or on time). Create a visualization to explore your hypothesis. Do the data appear to support your hypothesis?