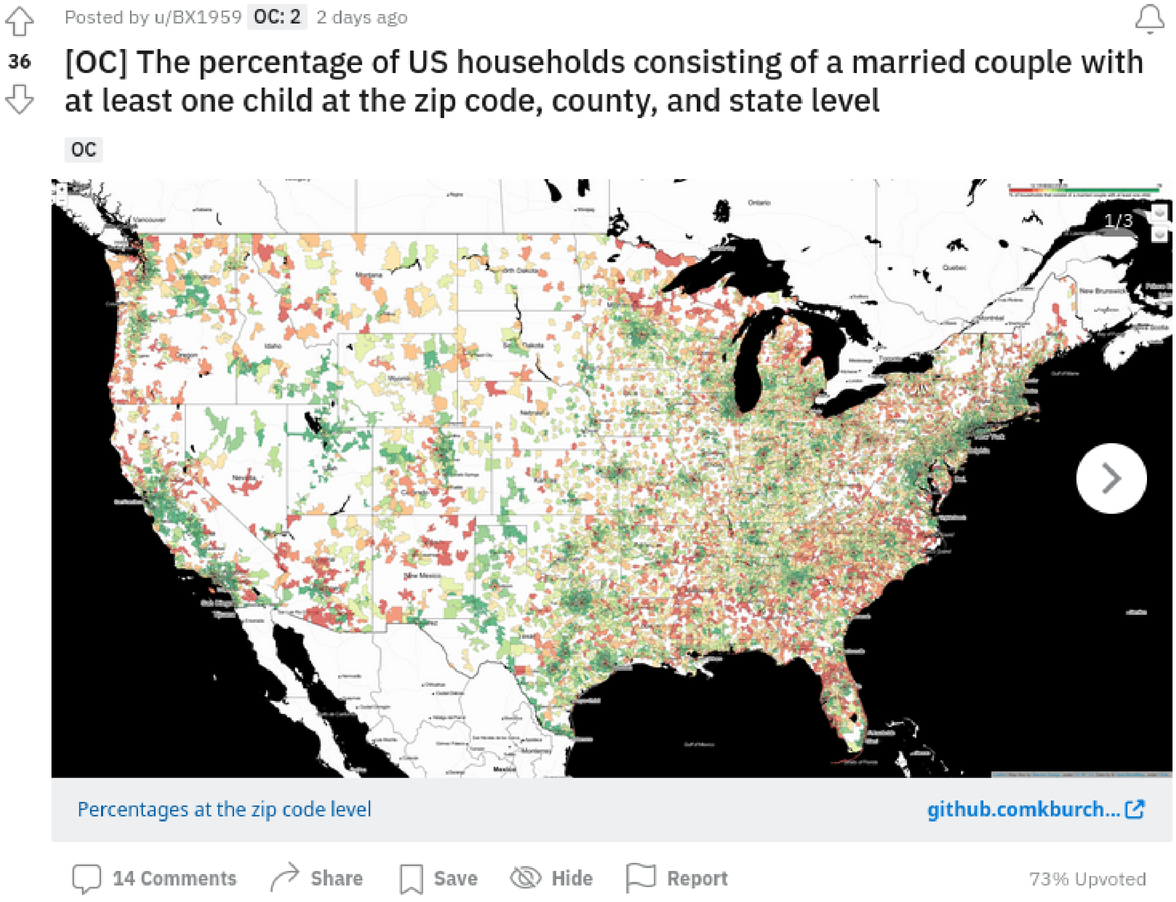
Data 100, Spring 2023

Exam Prep Section #3 Solutions

/r/datais[not so]beautiful

1. The following are 4 recent posts from the subreddit /r/dataisbeautiful. For each post, present at least one critique of the visualization, and at least one improvement or fix to address this critique.

(a)



Posted by u/BX1959 OC: 2 2 days ago

36

[OC] The percentage of US households consisting of a married couple with at least one child at the zip code, county, and state level

OC

Percentages at the zip code level

github.comkburch...

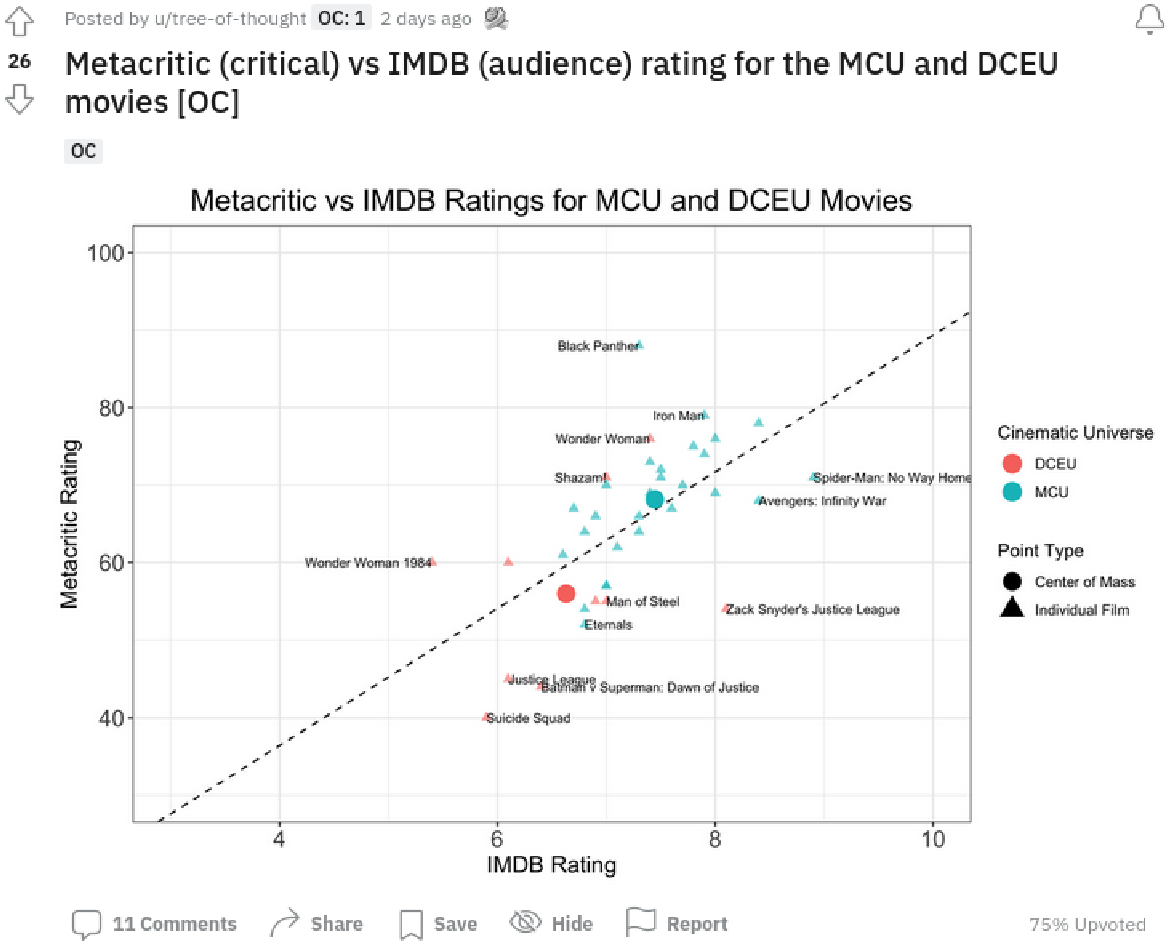
Solution: Here is one of Reddit's own critiques!

/u/BreqsCousin says:

I don't think I'd have chosen red/green for this. It's not super accessible. And it strongly suggests that green is good and red is bad.

Another critique is that the legend is far too small in relation to the plot. It's thus hard to see what threshold makes a percentage red vs yellow/green. Making the legend larger and using an alternative colormap (maybe one of constant color but variable opacity) that doesn't imply positive or negative connotations that may be misleading are a good set of fixes.

(b)



Posted by u/tree-of-thought OC: 1 2 days ago

26

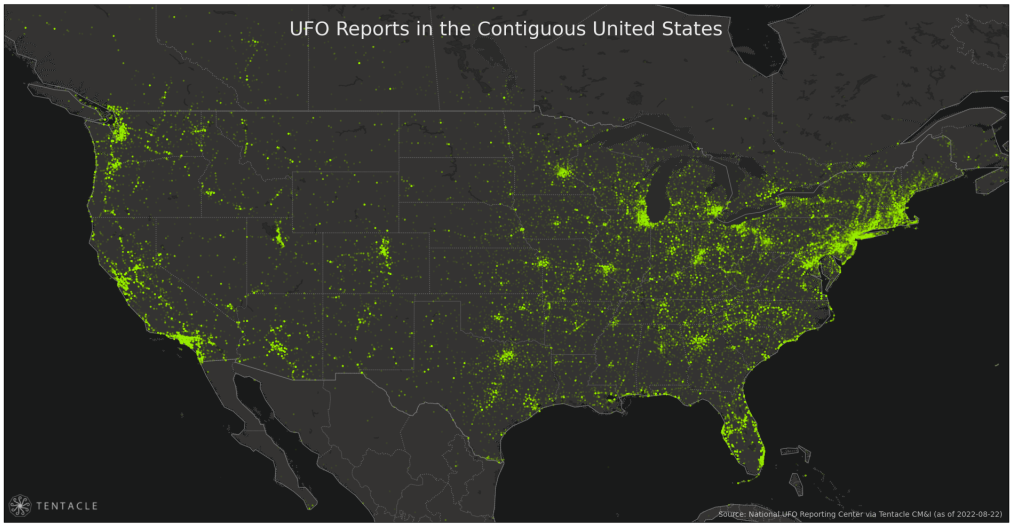
Metacritic (critical) vs IMDB (audience) rating for the MCU and DCEU movies [OC]

OC

Solution: Another Reddit criticism from /u/UsernameTaken1701:

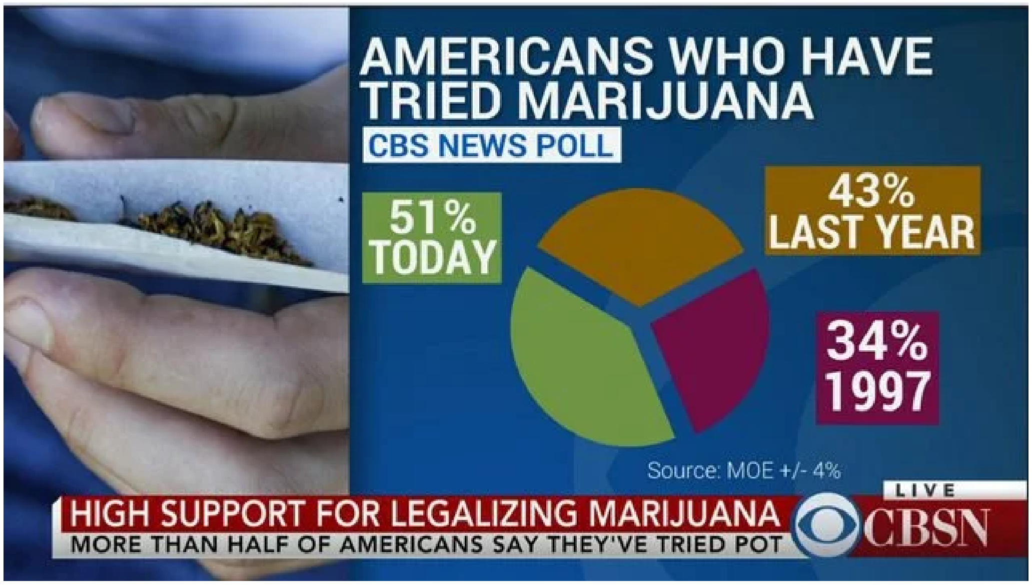
I can't rationalize this, but it bugs me Marvel is blue and DC is red. It feels like it should be the other way around.

There is some overlapping between points labels, and it is unclear how the line is relevant. Some of the movies have text labels and others don't, so it's unclear what these are supposed to convey. Personally, I would opt for no text labels at all and to remove the dotted line because it serves no purpose.


UFO Reports in the Contiguous United States
TENTACLE
Source: National UFO Reporting Center via Tentacle CM&I (as of 2022-08-22)

(c)

Solution: While the dark grayscale background sharply contrasts with the fluorescent green points (very cool!), there are a few issues here. First, we can probably assume brighter points mean more reports, but we don't know just how many. It would be nice to see a legend for this brightness axis. Additionally, if you zoom in, you'll see the brightness is not due multiple points being overlaid, but rather they are all single points that are just brighter. So, it's unclear if each point represents a single report, or a bunch of reports grouped in some geographical bin. Finally, this map isn't particularly informative: all it shows is that there are more UFO reports in big population centers, which seems like common sense. It would be interesting if they normalized by population of the various areas to see if some areas have a disproportionately high number of UFO reports. Definitely not worthy of the 11,000+ upvotes it actually got.



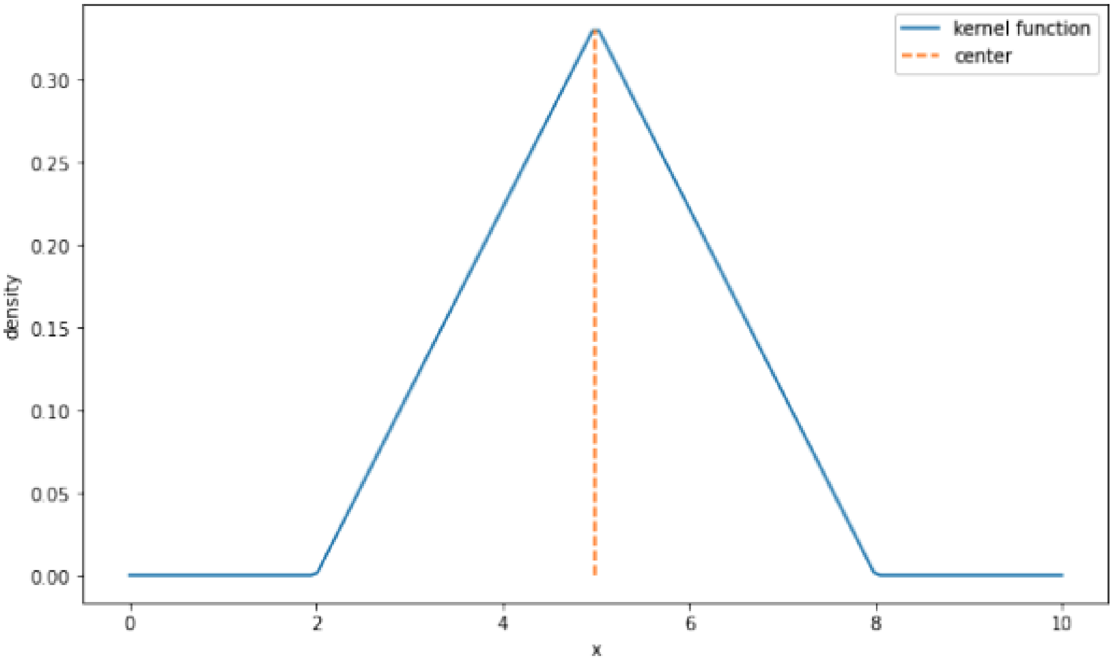
(d)

Solution: This image is actually from the counterpart subreddit of /r/dataisbeatiful, /r/dataisugly. Besides the attempt to use a pie chart to represent the distribution of a categorical variable, the data appears to not even be a single categorical variable. Rather, three different categorical variables are secretly being plotted on the same pie chart (each slice is the percentage of people that have tried marijuana in a survey at a particular time, so at each time the survey was conducted, we have a categorical variable with levels either ”yes” or ”no”). So, the percentages are not guaranteed to add to 100, and sure enough, they don’t. The source of the data is also the margin of error, which is strange. A more reasonable visualization might be stacked bar-charts for each time the survey was conducted, placed next to one another.

Triangular Kernel?

2. Brian finds that the boxcar and Gaussian kernels do not produce good visualizations for his dataset, so he decides to develop a new kind of (non-smooth) kernel called the triangular kernel! However, he is not sure how to finish the mathematical formulation for it! Luckily, he was able to generate a visualization that might help you help him.

   A plot of the triangular kernel centered at $z = 5$ with bandwidth $\alpha = 3$ is shown below.



Below is the incomplete mathematical formulation of the triangular kernel, where β is a variable for which he has not yet assigned a value in terms of the function inputs x and z and bandwidth parameter α.

$$T_\alpha(x,z) = \begin{cases} \beta - \frac{1}{\alpha^2}|x-z| & |x-z| \leq \alpha \\ 0 & |x-z| > \alpha \end{cases}$$

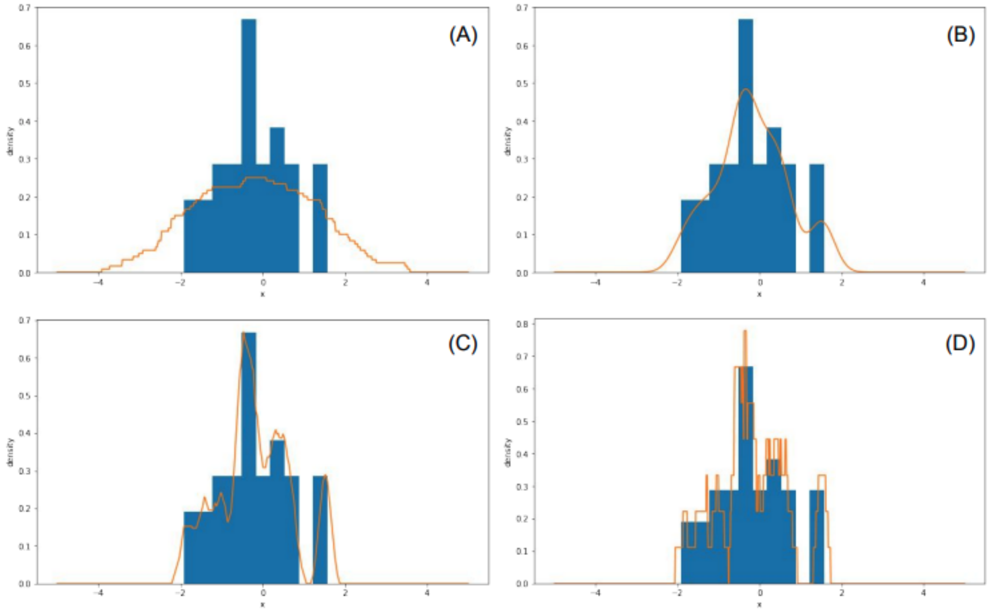
(a) What is the value of β, expressed as a function of α and/or z, that makes this a legitimate kernel to use in KDE? (Remember the area formula for triangles: $A = \frac{1}{2}bh$, where A is the area, b is the length of the base, and h is the height).

Solution: To be a legitimate kernel, the area under the triangle must be 1, so we wish to find β such that this area is always 1. From this function, we see that the triangle is centered at some value of z, representing some data point at which the kernel was placed. If we plug in $x = z$, which should give the height of the triangle, we obtain β. Note that the triangle function gives 0 for any x that is more than a distance of α away from z, and gives nonzero values for x that is within α of z. So, the base of the triangle is just the length of the interval for which x is within an α distance of z, which is 2α. So, $h = \beta$ and $b = 2\alpha$, so plugging into the equation for the area of a triangle, we obtain:

$$1 = \frac{1}{2}bh = \frac{1}{2}\beta 2\alpha = \beta\alpha$$

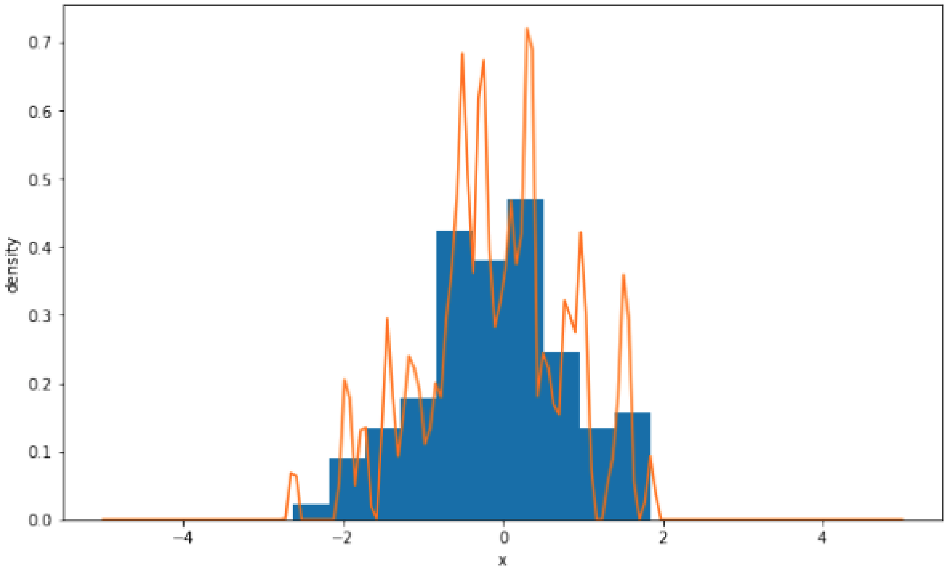
where we set the area to be 1 because this makes our kernel legitimate. Solving for β, we obtain $\beta = 1/\alpha$.

(b) Match the following plots with the most likely kernel functions out of the following options: Gaussian, boxcar, triangular. Clearly indicate which kernel you are assigning to which plot.



Solution:
(A) and (D) use boxcar kernels, since you can see the square edges from the individual kernels being placed at the data points. (B) is a Gaussian kernel since it's entirely smooth and you can see the classic bell-shaped curve at areas of high data concentration. (C) is the triangular kernel because you can see triangular peaks at areas of high data concentration.

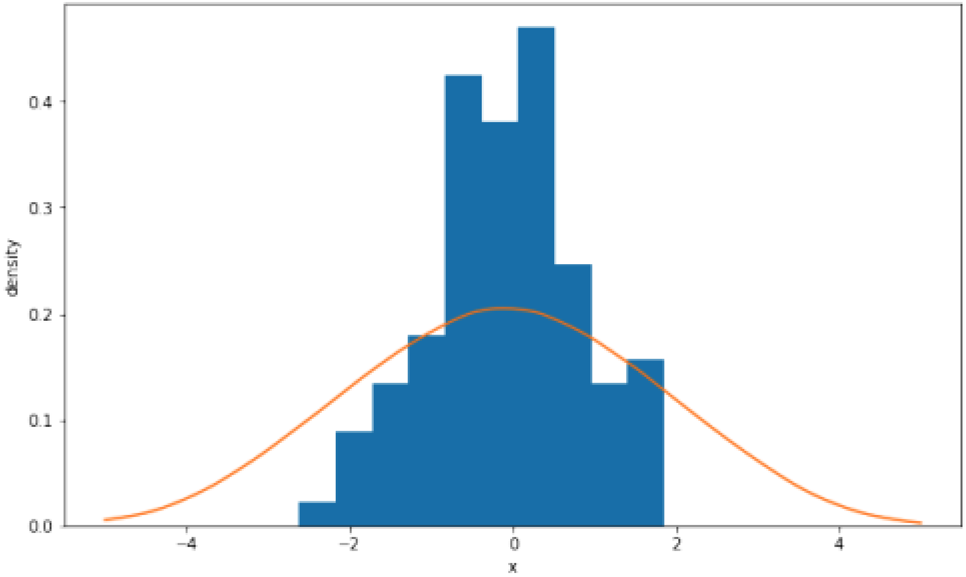
(c) Brian generates the following KDE plot using the triangular kernel. Is his choice of bandwidth parameter α too large or too small? Explain your reasoning in terms of the provided function for the triangular kernel.



Solution:

By the function provided for the triangular kernel, α controls how wide the base of the triangle is (higher alpha means wider base). Since we found in part (a) that α is inversely proportional to the height, increasing alpha also makes the triangle shorter and more stubby. α is much too small here since we observe a lot of thin spikes.

(d) He tries another value of α to correct for the issue identified in the last part, and ends up with the following plot. Is his new choice of α too large or too small? Explain your reasoning in terms of the provided function for the triangular kernel.



Solution:

Using the same assessment of α on the shape of the kernel from the previous question, we can conclude that α is too large since the resulting KDE is very wide and short. This is consistent with how we expect the individual kernels to behave with large α.