DS 100: Principles and Techniques of Data Science **Date: February 22, 2019**

Discussion #5

Name:

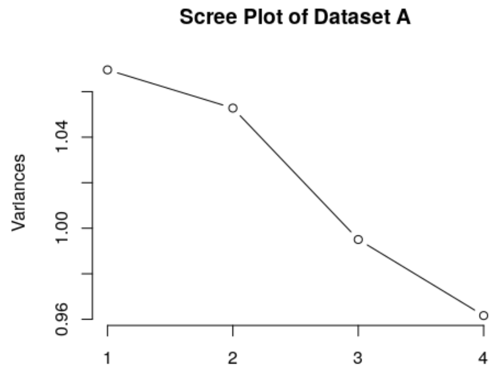
Dimensionality Reduction

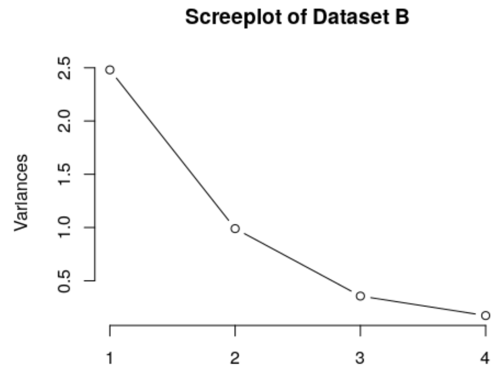
1. Principal Component Analysis (PCA) is one of the most popular dimensionality reduction techniques because it is relatively easy to compute and its output is interpretable. To get a better understanding of what PCA is doing to a dataset, let's imagine applying it to points contained within this surfboard. The origin is in the center of the board, and each point within the board has three attributes: how far (in inches) along the board's length, width, and thickness the point is from the center. These three dimensions determine the spread of the data.

   (a) If we were to apply PCA to the surfboard, what would the first three principal components (PCs) represent? Feel free to draw and label these dimensions on the image of the surfboard.

   (b) Which of the three PCs should be used to create a 2D representation of the surfboard? How come? Make a sketch of the 2D projection below.

2. Compare the scree plots produced by performing PCA on dataset A and on dataset B. For which of the datasets would PCA provide a scatter plot that describes the variability of the data without leaving out much information? Note that the columns of both datasets were centered to have means of 0 and scaled to have a variance of 1.





Midterm Review

1. Probability and Sampling

3. A small town has 5 houses with the following people living in each house:

Suppose we take a **cluster sample** of 2 houses (without replacement), what is the chance that:

(a) Kim and Lars are in the sample

○ 0 ○ 1/20 ○ 1/10 ○ 1/6 ○ 1/5 ○ 2/5 ○ 1

You may show your work in the following box for partial credit:

(b) Kim, Abe, and Ben are in the sample

○ 0 ○ 1/20 ○ 1/10 ○ 1/6 ○ 1/5 ○ 2/5 ○ 1

You may show your work in the following box for partial credit:

(c) Kim and Dan are in the sample - **Select all that apply**

☐ The same as the chance Kim and Lars are in the sample

☐ The same as the chance Kim, Abe, and Ben are in the sample

☐ Neither of the above

2. Transformations and Smoothing

4. Which of the following are reasonable motivations for applying a power transformation? **Select all that apply**:

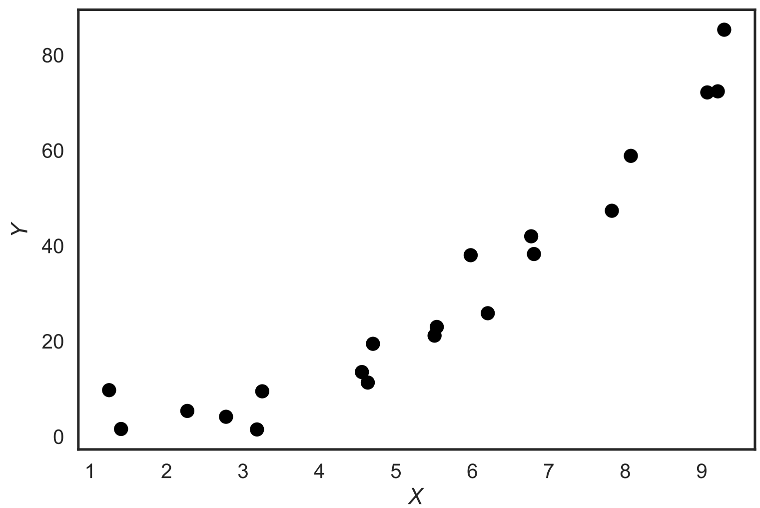
   ☐ To help visualize highly skewed distributions

   ☐ Bring data distribution closer to random sampling

   ☐ To help straighten relationships between pairs of variables.

   ☐ Reduce the dimension of data

   ☐ Remove missing values



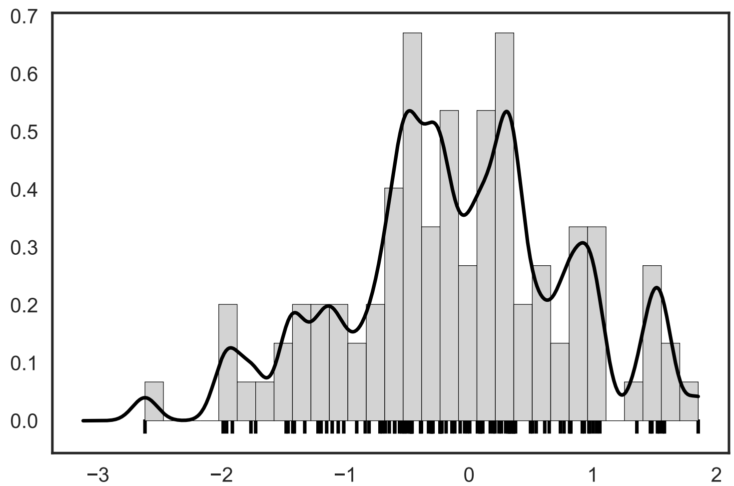
5. Which of the following transformations could help make linear the relationship shown in the plot below? **Select all that apply**:

☐ $\log(y)$ ☐ x^2 ☐ $\sqrt{y}$ ☐ $\log(x)$ ☐ y^2 ☐ None of the above

6. The above plot contains a histogram, rug plot, and Gaussian kernel density estimator. The Gaussian kernel is defined by:

$$K_\alpha(x, z) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{(x-z)^2}{2\alpha^2}\right)$$

Judging from the shape of separate standing peaks, which of the following is the most likely value for the kernel parameter α.



○ $\alpha = 0$ ○ $\alpha = 0.1$ ○ $\alpha = 10$ ○ $\alpha = 100$