

DATA 100: Vitamin 6 Solutions

September 27, 2019

1 Visualizations: Choosing the Scale

When plotting data, which of the following approaches to axis scaling should be avoided?

- ☒ Using different scales for variables on the same axis
- ☒ Changing the scale mid-axis
- ☐ Standardizing the scales for the same axis
- ☐ Maintaining a consistent scale through the axis

Explanation: Scales should be standardized when plotted on the same axis. This insures that trends and/or relationships in the data are not distorted. Scales should remain static. Changing the scales along an axis makes for difficult to interpret plots.

2 Revealing the Data

Which of these statements should be followed to create effective visualizations?

- ☐ Each datapoint should have a unique color
- ☒ Avoid over-plotting
- ☒ Avoid spurious dimensions
- ☐ Avoid using legends and axes labels

Explanation: It is best to avoid visual noise when plotting, such superfluously coloring each observation in the plot. For this reason, it is also best to avoid over-plotting and the inclusion of spurious dimensions. Though we wish to reduce visual clutter, legends and axes labels play important roles when interpreting a plot. They should always be included.

3 Canonicalization

Fill in the blank: Canonicalization consists of converting data that has more than one possible presentation into ----.

- ☐ a DataFrame.
- ☒ a standard form.
- ☐ a canon.

4 Regular Expressions

The language of SSNs is described by which regular expression?

- ☐ "[0 to 9]3-[0 to 9]2-[0 to 9]4"
- ☐ "0-9[3]-0-9[2]-0-9[4]"
- ☒ "[0-9]3-[0-9]2-[0-9]4"
- ☐ "[0-9]3-[0-9]3-[0-9]3"

Explanation: See slides for details.

5 Regular Expression Syntax

Which of the following strings would match with "jo*hn"?

- X joohn
- ☐ jon
- ☒ jhn
- ☒ john
- ☐ jooooooooohnn

Explanation: Recall that "o*" will match with a string containing zero or more "o"s.