HW 1:
EDA and Visualization

Prof: Yannet Interian

Due: September 1st 2020 11:59 PT

Submit this assignment on canvas as a pdf.

1 Cereal data analysis and Storyboard (4 points)

Readings for this problem:

- "Storytelling with data" Chapter 1.
- "Storytelling with Data: Let's Practice" Exercise 1.3, 1.7, 1.8, and their respective solutions.

You work for a large school district as a data analyst. Your boss wants to purchase a large amount of cereal for school breakfasts. He needs to choose a manufacturer and product. He wants you to prepare a presentation for the executive team. You are provided with some data.

You are allowed to load your data in pandas and look at column names, and data to understand column types and context, but **don't use any visualization tools** (no matplotlib). Follow the recommendations from the books and the lectures. Respond to the following questions:

1. List five analytics queries or questions that you would have about this dataset in your exploratory process.

2. Describe 5 visualizations that can help you explore this data. (Example: line graph between variable x and y).

3. What other data would be helpful to prepare your presentation? What do you think is your boss's goal?

4. Assume that after a long exploratory data analysis you reach a recommendation for your boss (make assumptions as needed). Provide a one-sentence "big idea."

5. Create a storyboard for your presentation.

We will use the following questions to guide our grading:

- Are the questions in 1 related to the manager's ask?
- Are the visualizations following the advice/ guidelines given in class?
- Does 3 make sense?
- Does the "big idea" have the 3 components: (1) articulate a point of view; (2) convey what is at stake; and (3) be a complete sentence?
- Does your storyboard make sense? Is it related to your "big idea"?

Dataset: "cereal.csv" .The data should be pretty self-explanatory. The Manufacturer is a one-letter code with the mapping: Q-Quaker Oats, P-Post, G-General Mills, K-Kelloggs, R-Ralston Purina, N-Nabisco. Type stands for C (cold) or H (hot). Shelf stands for which row on a shelf the cereal is on (1=bottom, 3=top). The rest are attributes that describe the nutritional contents of the cereal.

2 Practice basic matplotlib (3 points)

Readings for this problem:

- "Storytelling with data" Chapter 2
- "Storytelling with Data: Let's Practice" Chapter 2.

Practice using matplotlib to plot the mortality rate over time of children under 5 (per 1000 live births). Submit 3 visualizations for this data. Which one do you think works the best? Submit a pdf of your code with your plots. Make sure your plots have the following features: (Taken from DataVizChecklist by Stephanie Evergreen & Ann K. Emery)

- 6-12 word descriptive title is left-justified in the upper left corner
- Text size is hierarchical and readable
- Text is horizontal
- Proportions are accurate
- Data are intentionally ordered
- Color is used to highlight key patterns
- Gridlines, if present, are muted

Dataset: "usa_mortality_rates.csv"

3 Practice slopgraph with matplotlib (3 points)

Draw the plot below in matplotlib. Do your best to get as close as you can.

Hints: Note the different shades of grey; the alignment of the text; the thickness of the lines. One way to do this is to draw segments; then add points and text.

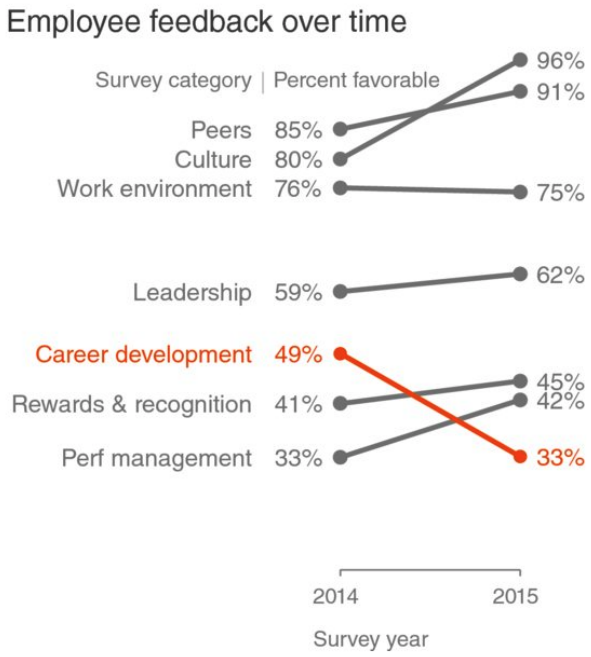

Figure 1: Slopegraph from "Storytelling with data"