Problem Set 1
Part 2: Critical Thinking

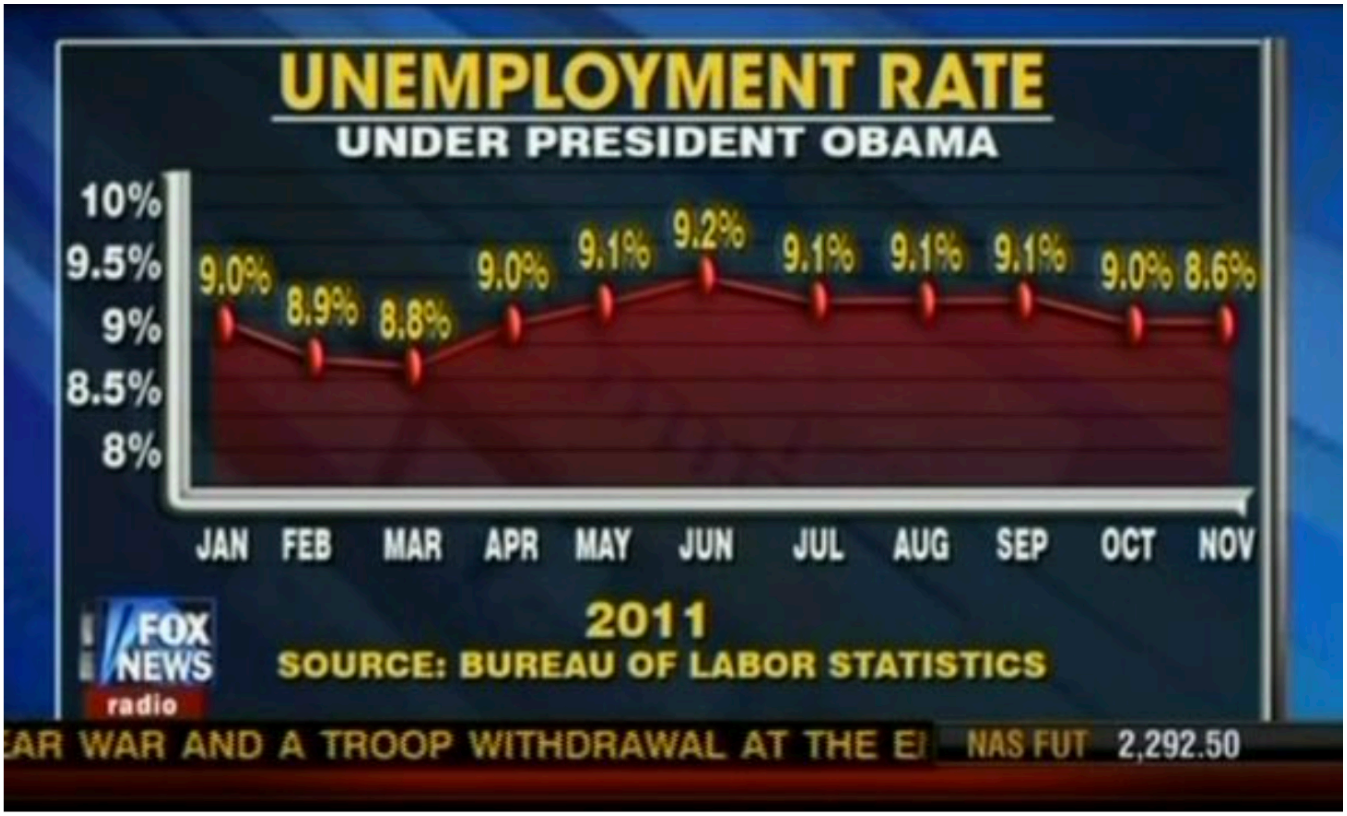
Di Tong

1. Describe the different information contained in/revealed by visual versus numeric exploratory data analysis. (*Hint*: Think of different examples of each and then what we might be looking for when leveraging a given technique).

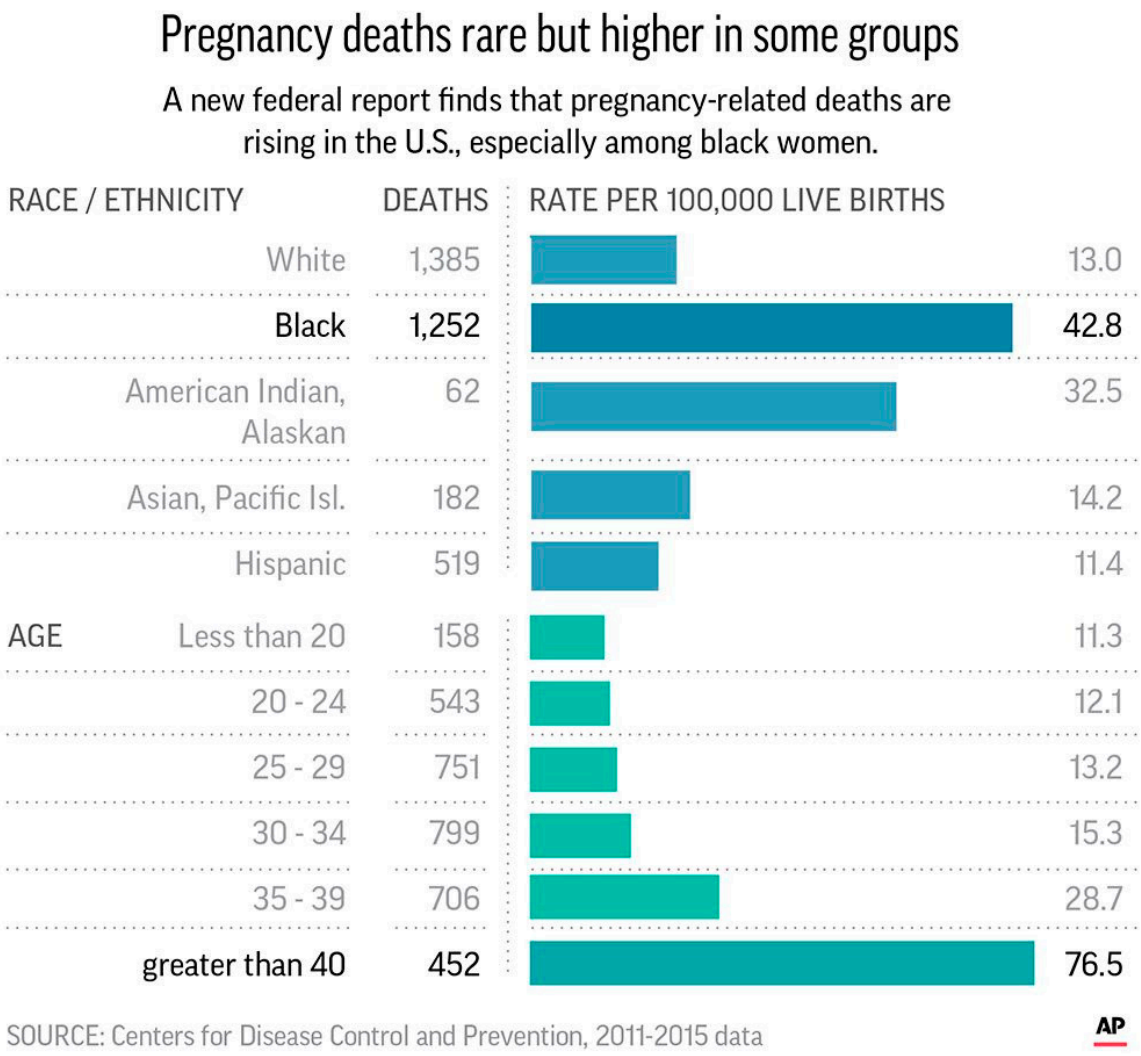
Visual exploratory data analysis can more intuitively and directly demonstrate interesting patterns that could inform important research questions, directions and findings. It is also easier for researchers to have a sense of the entire picture, and hence easier for detecting outliers and relationships.

Numeric exploratory data analysis is more exact and specific in delivering information. Correspondingly, it usually contains less information as a tradeoff with its exactness. Owing to its relatively higher level of abstraction, information regarding anomalies would be overlooked or even hamper the accuracy of the numeric description.

2. Find (and include) two examples of "bad" visualizations and tell me precisely why they're bad.



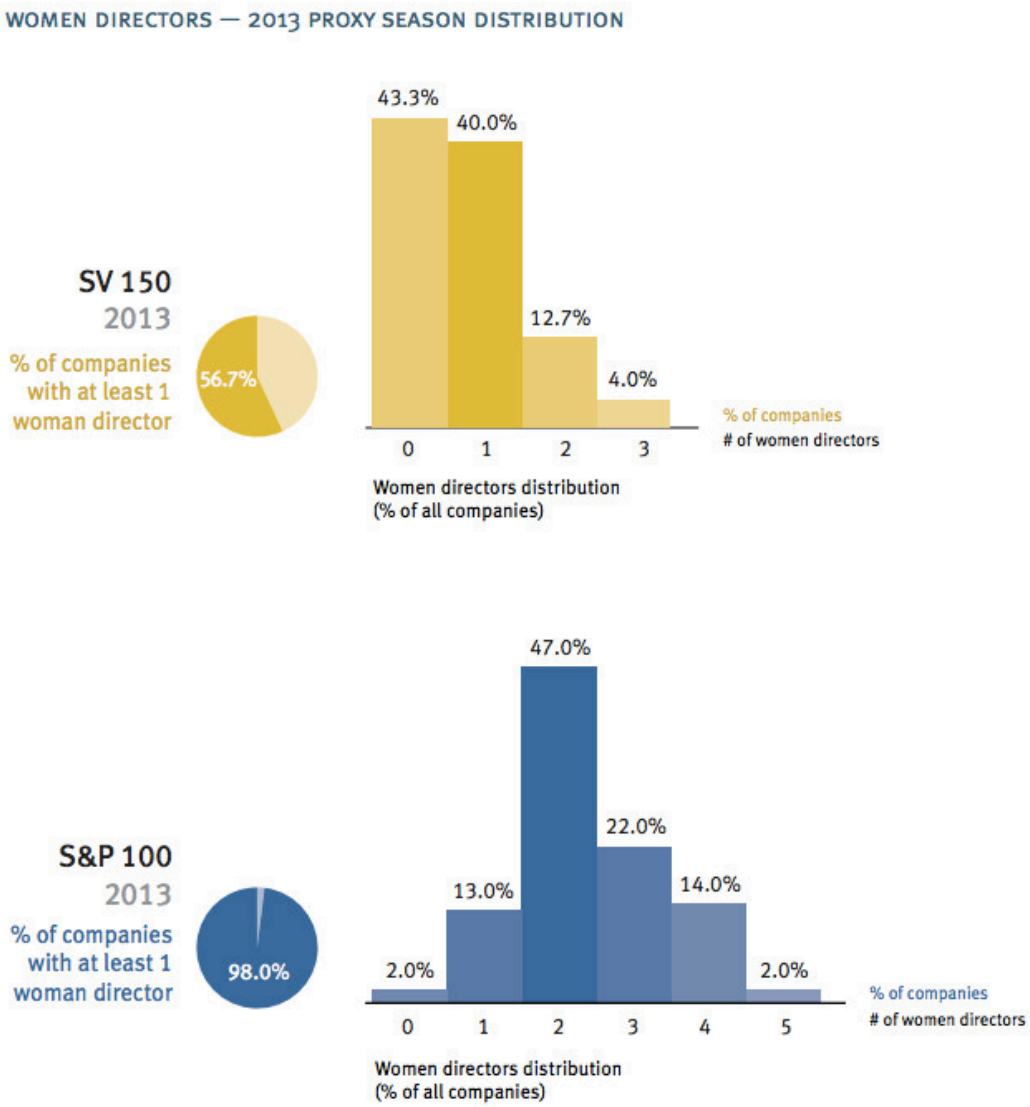
This major problem for this visualization lies in its truthfulness. The 8.6% at November looks higher than the 8.8% in March. In addition, as the y-axis ends at 10%, the unemployment rate appears to be high in this graph, while it is only around 9%. What's more, the y-axis starts from 8% and ends at 10%, making the yearly fluctuations between only 0.6% looks more significant than they actually are.



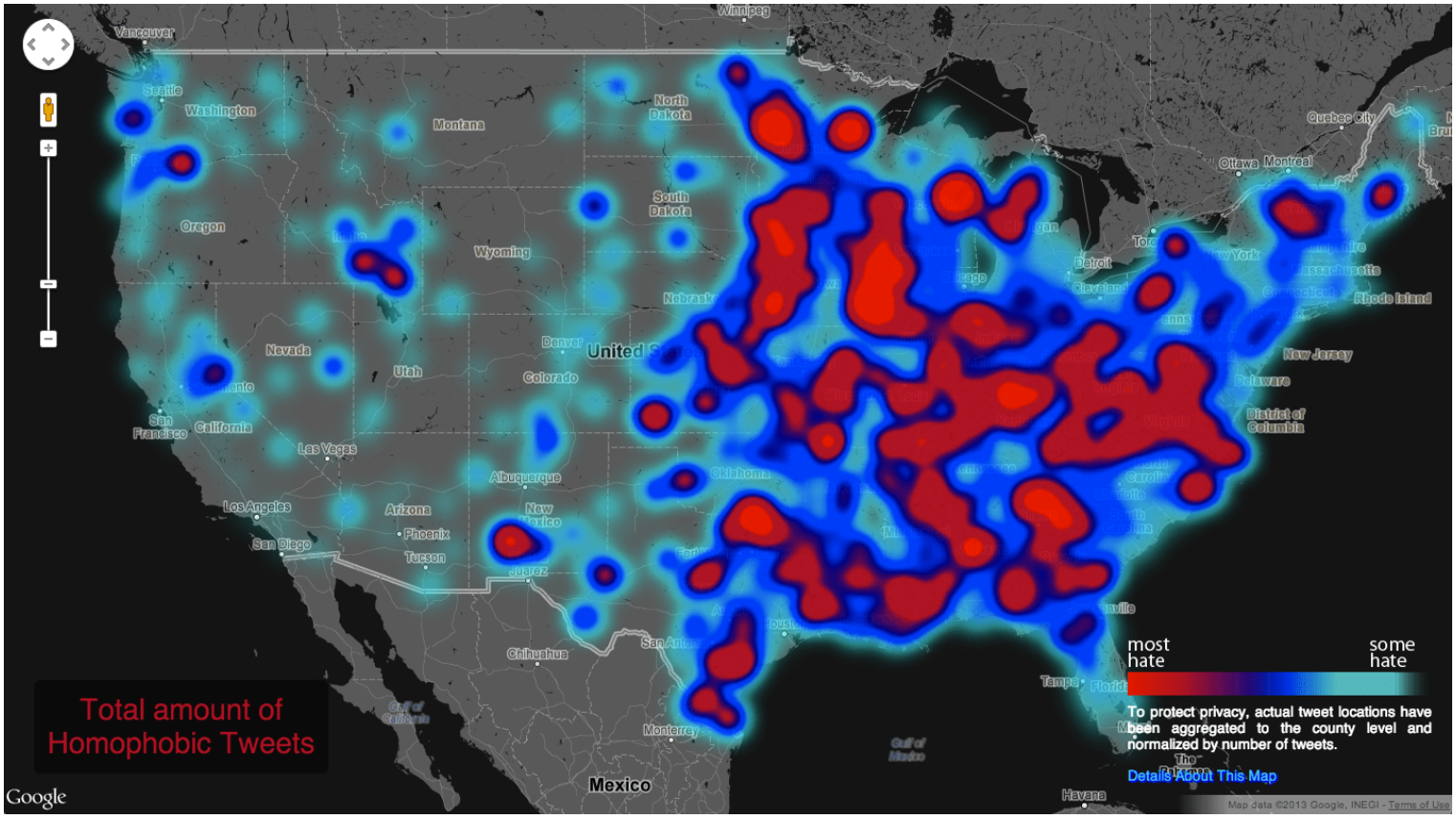
This visualization has several problems. First, the length of the bar (greater than 40) is similar to that of the bar (Black), while the rates are 76.5% and 42.8%, respectively. Second, the visualization looks messy, as the order of categories is not sorted according to the pregnancy death rate or number. It is also unnecessary to use two different colors for the same variable. Besides, though the title mentioned the rising trend of pregnancy death rate, it is not demonstrated in this graph.

3. Find (and include) two examples of "good" visualizations and tell me precisely why they're good.

Title: The distribution of women on the boards of directors at companies in the Silicon Valley 150 (the Valley's biggest companies by sales), compared to the boards of the 100 biggest public companies in the U.S. (Standard & Poor's 100)



This group of visualization is good for it is concise, clear, beautiful and enlightening. There's no unnecessary information. The juxtaposition of the two women director distributions and the two piecharts clearly and vividly illustrate the story of severe gender bias in tech companies. Besides, the whole format and style is consistent and beautiful.



This graph does a good job in elegantly demonstrate the regional variation in homophobia in the U.S. It is very clear and easy to interpret even though there's not much captions and labels presented. The use of color also corresponds to human emotion as well as the political ideology related to homophobia. Besides, the caption under the legend explains the measurement used for this visualization, making it more informative to the audience and hence enhancing its truthfulness.

4. When might we use EDA and why/how does it help the research process?

We should use EDA when we want to detect and formulating meaningful and important questions, guide the research design, data collection and analysis. EDA can facilitate the implementation of the very confirmatory paradigm. It is especially helpful in terms of identifying questions and directions for research.

5. What did John Tukey mean by "confirmatory" versus "exploratory"? Give me an example for each.

"Confirmatory" analysis denotes a research paradigm that aims at and operates in the way of testing given hypotheses. Example: examining the relationship between automation level and

employment rate in the automobile manufacturing industry with a hypothesis that automation development would bring about more unemployment in the automobile manufacturing industry.

“Exploratory” analysis refers to efforts that seek to find the research questions, directions, hints of explanations in a data-driven way without hypotheses. It is “an attitude, a flexibility and a reliance on display.” Example: apply dynamic topic modeling on the UN conference report during the last century to see the temporal pattern regarding UN’s major concern