# Behavioral Research and Experimental Design
# Assignment 3

Deadline: 8 November, 2023 (2359 Hours)

1. Examine the shortcomings in the provided visualizations, describe the initial impressions readers might form when viewing them, clarify the actual data representation they convey, and propose alternative visualization methods. [2M]

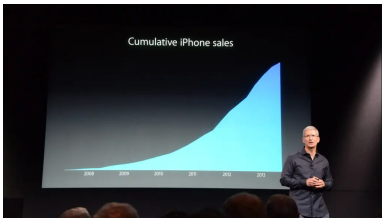


Figure 1: Tim what's cooking?

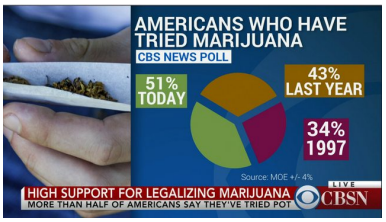


Figure 2: Weeding out them stats!

2. The Wine Quality Dataset contains information about different types of wine and their associated quality ratings. It includes various chemical properties of the wines, such as alcohol content, acidity, and sulfur dioxide levels. You are tasked with analyzing various chemical properties of wines and their quality ratings. Select one chemical property (e.g., alcohol content, acidity, sulfur dioxide levels) and calculate the mean, median, mode, and standard deviation for that property. Next, create a visualization to detect outliers for the chosen property. Identify and list the outliers based on the visualization.

   Afterwards, remove the outliers from the dataset for the chosen property and recalculate the mean, median, mode, and standard deviation. To deepen your analysis, repeat these steps for multiple chemical properties. Discuss the differences in descriptive statistics across these properties before and after outlier removal. What do these variations suggest about the influence of outliers on different descriptive statistics as well as aspects of the data? Elaborate on the importance of outlier detection in your analysis. [5M]

3. Provided with the Gapminder Dataset, which contains data about countries' life expectancy and GDP per capita over time.

   (a) Your task is to create an inappropriate visualization to represent GDP per capita over time for India. Explain why this visualization is not suitable for representing the provided data. [2M]

   (b) Create an appropriate visualization and argue why it is best suited for the current scenario. [1M]

4. You are provided with a dataset containing a single column named "Value" in the form of a CSV file. Your task is to assess whether the data in this column is normally distributed. If not, identify the distribution and apply a suitable transformation to make the data more representative.

   (a) Explain how you assess normality, identify the data's distribution, and describe any visualizations or tests used. [1M]

   (b) If the data isn't normally distributed, specify the observed distribution and its characteristics. [1M]

   (c) Apply an appropriate transformation and clarify why you chose it. [1M]

   (d) Select and compute the most suitable descriptive statistics before and after transformation, justifying your choices and explaining how they change with the transformation. [2M]