Lecture 1 – Introduction and EDA Fundamentals

Motivation:

1. Salaries
    a. Statistician $80,110 per year base median
    b. Actuary $97,070 per year base median
    c. Data Scientist $120,931 per year base average

2. Statistics as an Art - not as black and white as math
    a. According to the Motor Vehicle Website, the % of adults who report driving after drinking too much in the last 30 days is as follows:

Two Points to consider:

i.

ii.

Statistics: the study of the collection, organization, and interpretation of data. Our job is to:

i.

ii.

BIG picture:

Some Terminology:

1. **Population** – entire collection of people or objects of interest
2. **Sample** – subset of the population for which we actually obtain data
3. **Inference** – drawing conclusions about the population based on the data collected from the sample
4. **Parameter** – a number summarizing some feature of the population
    a. Ex: average, proportion
5. **Statistic** – computed solely from the data (from the sample) and summarizes some corresponding feature of the sample

Several Approaches to Statistics

1. Classical Approach

2. Bayesian Approach

3. EDA Approach

The Statistics Lifecycle:

Notes on the Statistics Lifecycle:

1) Typically, EDA is graphical and descriptive (uses ALL the data)

2) The KEY – no assumptions placed on the data vs the classical approach

3) EDA – the goal is to understand the data
    a. Yes, it is important to understand what is in the data
    b. But it's almost MORE important to understand what is NOT in the data

EDA – Fundamentals

Terminology

1. **Data** – information regarding specific experimental units/subjects (usually individuals) organized in tabular form with certain variables

2. **Experimental Units/Subjects** – the people or objects for which we have information on (described in the dataset)

3. **Variables** – a characteristic that varies from subject to subject; usually identified by a column in a spreadsheet

Types of Variables: Quantitative vs Qualitative

1. Quantitative Variables – characteristics that can be measured
    a. Discrete –
    b. Continuous –
2. Qualitative Variables – characteristics observed
    a. Nominal –
    b. Ordinal –

Structure of a Dataset

1. Y variable = **Response Variable**, Variable of Interest, Dependent Variable
    a. For this course there is only 1, but there are methods to analyze more than one
2. X variables = **Covariates**, Independent Variables, Explanatory Variables

Examining Distributions – One Variable

One Categorical Variable

1. First need table of counts and percentages

a. Ex: Body Image What is your perception of your own body? Do you feel that you are overweight, underweight, or about right? A random sample of 1200 U.S. college students were asked this question as part of a larger survey.

Category	**Count** (**Frequency**)	**Percent** (**Relative Frequency**)
Overwight		
Underweight		
About Right		
Total		

2. Visualize using either a Pie Chart or a Bar Chart

3. Characteristics of a Pie Chart

a. Each piece of the pie is created using the frequency and relative frequency of each level of the variable.

b. Frequency → the number of subjects that fall into the specific category

c. Relative Frequency → the percent of all subjects that fall into the category

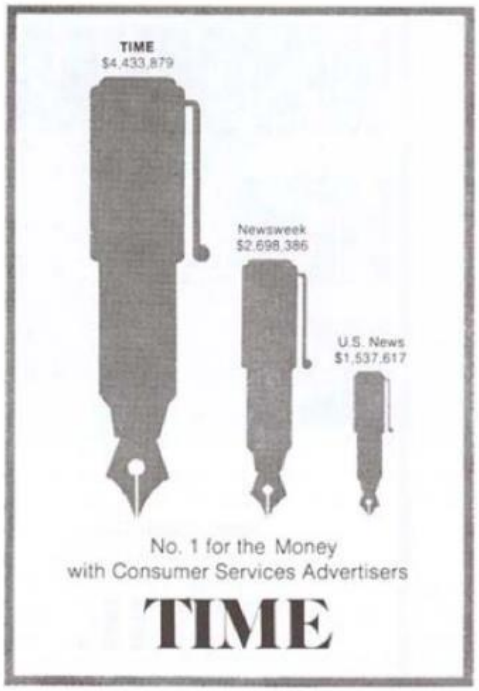
4. Characteristics of a bar graph

a. The sum of the bars is equal to the total number of subjects

b. Different types of bar graphs (Simple, Cluster, Stacked)

5. WARNING: Graphs can be misleading

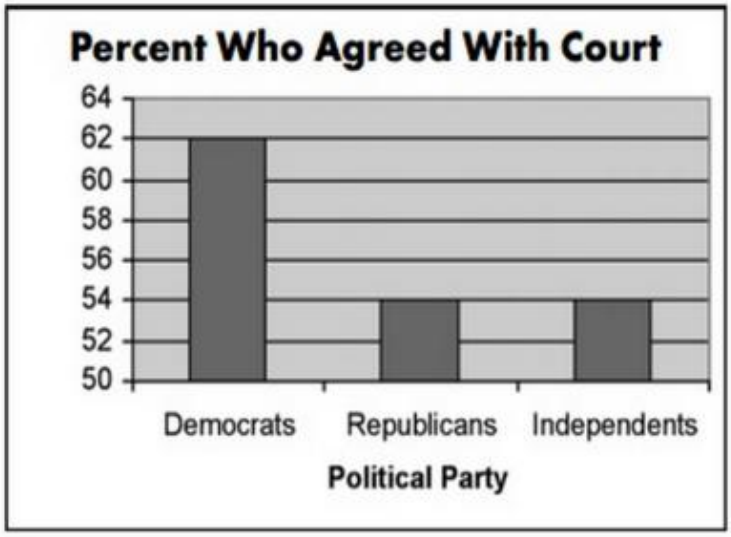
a. Violations of the **Area Principle –**



What's wrong with this?

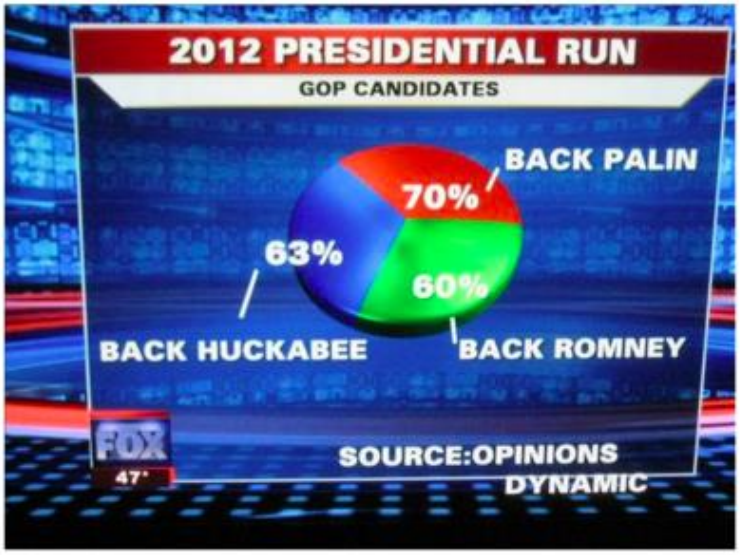
b. Be VERY careful of the **scaling**

i. Ex: Terry Schiavo was removed from life support after a years-long court battle. CNN used a graph similar to the one below to show who agreed with the decision to remove the feeding tube.



What's wrong with this?

c. Be careful of just poorly crafted graphs. Fox News is just the best at statistics...



I'm not even going to ask…

To summarize:

1 Categorical Variable:

Graphical:

Numerical Summary: