ECONOMICS 150:
Quantitative Methods for Economics

Class 2

- More on Summary Statistics
- Data Visualization

Measures of location

- What do we mean by location?
- There is more than one notion:
  - Mean: simple average
  - Median: mid-point (half the data is smaller, and half is larger)
  - Quartiles (percentiles): 1st quartile is the median of the bottom half, etc.
  - Mode: most common observation

Excel Functions:

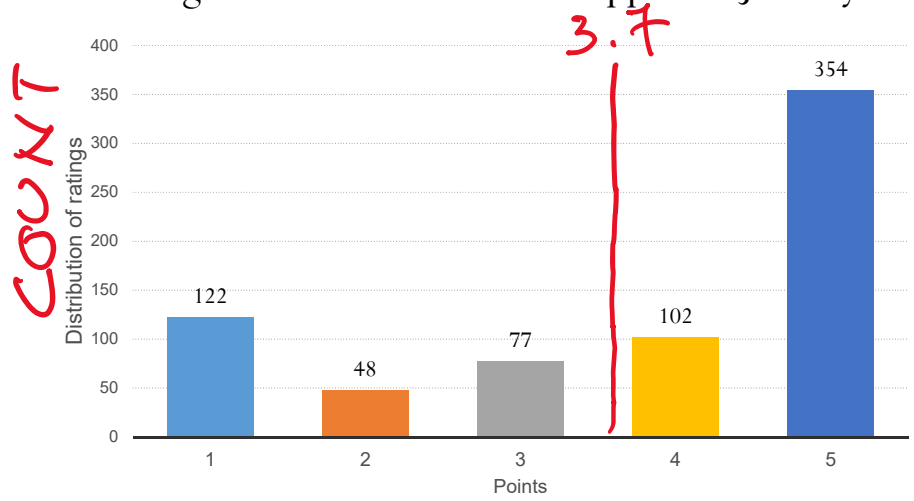
- AVERAGE(**array**)
- MEDIAN(**array**)
- QUARTILE(**array,num**)
- MODE(**array**)

- Which one is better?

Measure of location

MODE? 5 MEDIAN? 5 MEAN? 3.7

- Customer ratings for PAT track Android app as of January 2022



Measures of Location

- These tend to answer the question: where is the data?
  - Mean: add all observations and then divide the total by the number of observations:

the mean of $x_1, x_2, \ldots, x_N$ is $\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$

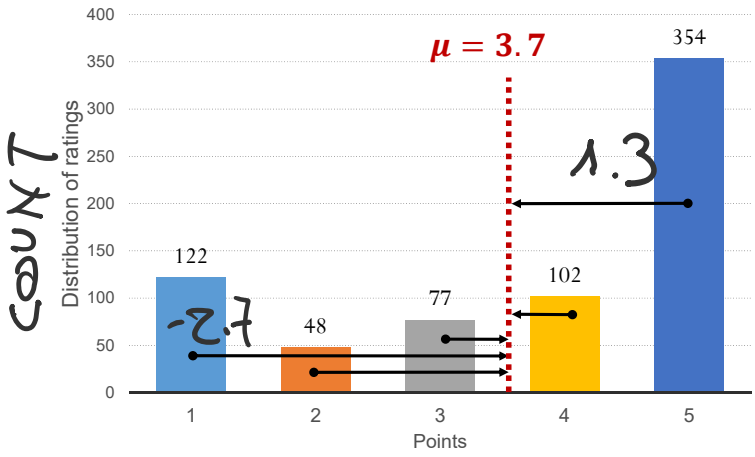
  - Median: is the value separating the higher half of a set of data values, from the lower half.
  - Mode: is the value that appears most often.

Measures of dispersion

- What do we mean by dispersion?

- Dispersion, Variability, Spread, "Risk"
  - Range — MAX(**array**) - MIN(**array**)
  - Variance — VAR(**array**)
  - Standard Deviation — STDEV(**array**)

Measures of dispersion with respect to the Mean



The variability is measured via deviations of the realizations from the mean

- Measure of spread (dispersion):
  - take each deviation into account *(potentially emphasizing big deviations)* by squaring them and then count how common each deviation is.

Measures of Dispersion

- These tend to answer the question: how spread out is the data?
  - Variance: is calculated summing all the squared difference of each data point from the mean, and then dividing the total by the number of points:

$$\text{the variance of } x_1, x_2, \dots, x_N \text{ is } \quad \sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

  - Standard Deviation: is the square root of the variance: $\sigma = \sqrt{\sigma^2}$
    - Why take the square root? Because it yields the original units.
  - Range: is the difference between the largest and smallest value in the data.

Customer Demographics

A company sold 117 of its signature items last year. The ages of 117 customers who bought these items are recorded in the file CustomerDemographics.xlsx. What are the mean, the median, the mode, the variance, and the standard deviation for these customers?

We can figure out the answer by going to Excel: CustomerDemographics.xlsx.

What do all these numbers mean?

Top Hat Attendance Code for Today

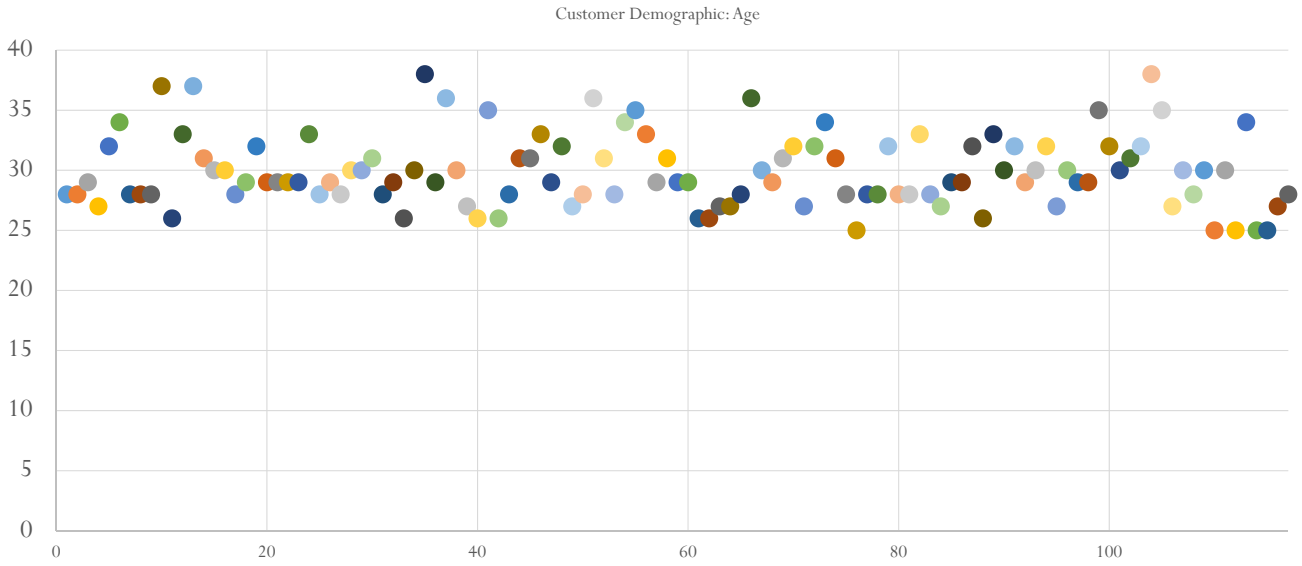
5323

Statistical Graphics

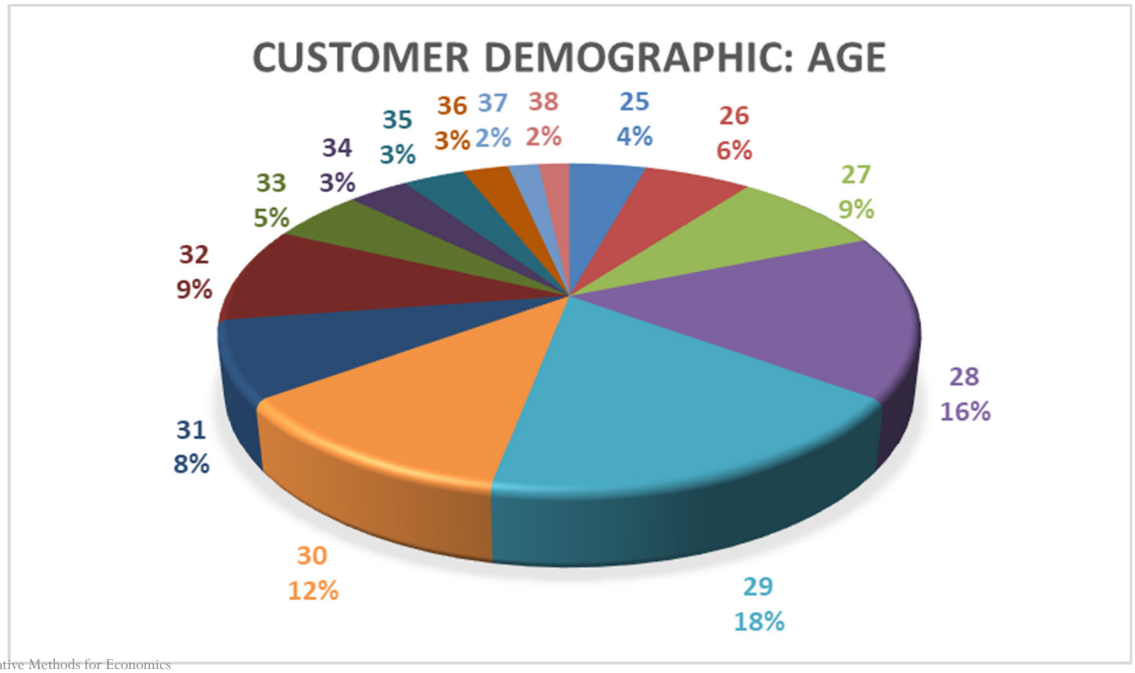
- Some Basic Ways to Summarize Data Graphically for a Single Variable
  - Pie Chart
  - Bar Chart
  - Histogram

- For two variables, one can plot relationships using scatterplots
  - These will be used a lot in the second half of the course

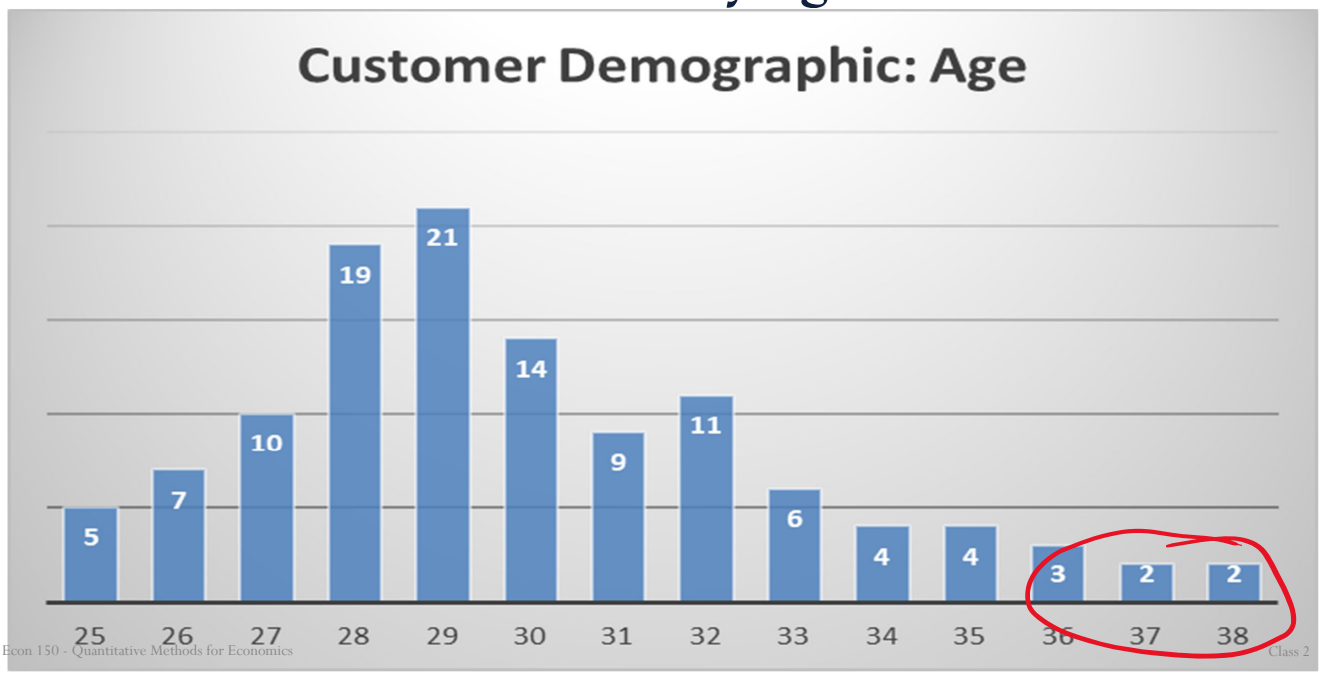
Picture of the Data: Customer Demographic



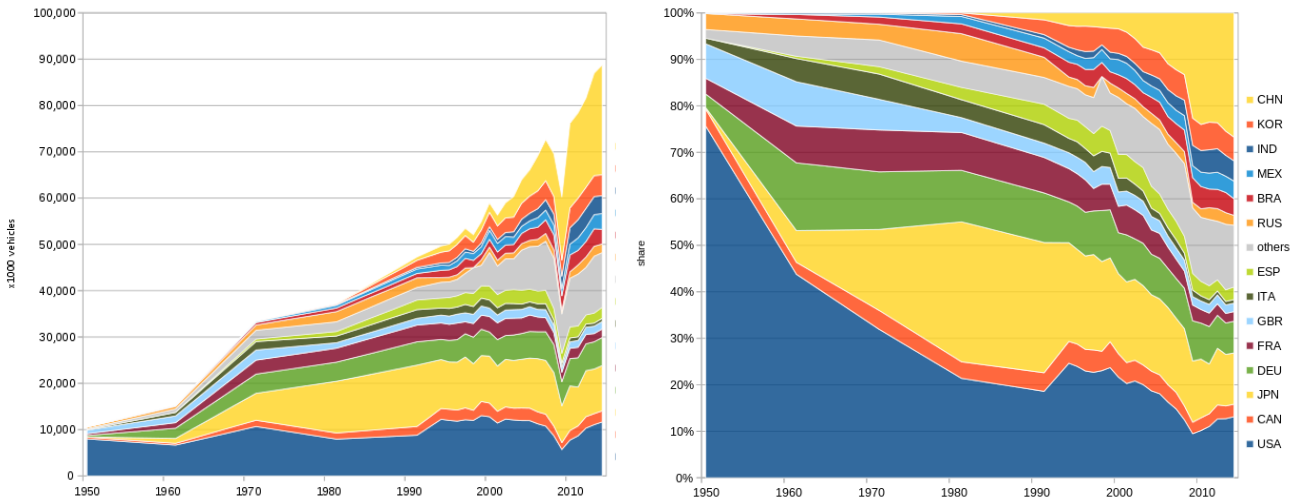
Pie Chart



Bar Chart by Age



Car Production by Country and Year



- Graphs can convey different messages depending on how one presents data

Data Visualization

- How does one think about visualizing data?
  - Objectives
  - Bad graphs / Good graphs
  - Tufte's principles

Communicating Data Through Images

- Data can be described not just by numbers and tables (which are useful), but also by pictures.

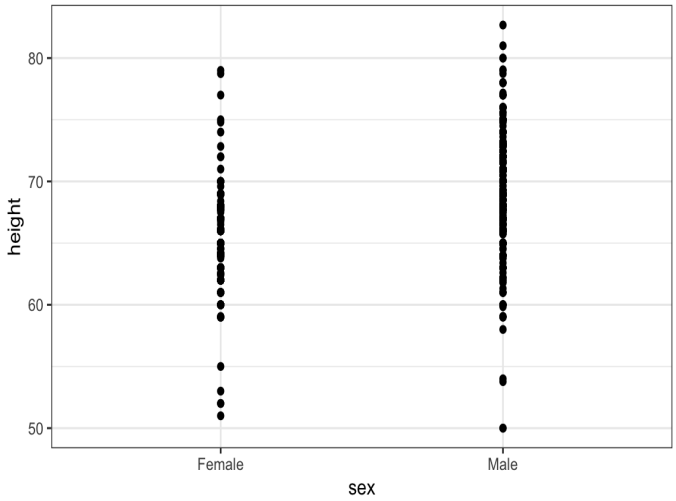
- Graphics display quantities by the combined use of **points**, **lines**, a **coordinate system**, **numbers**, **symbols**, **words**, **shading**, and **colors**

- At their best, graphics are instruments for reasoning about quantitative information

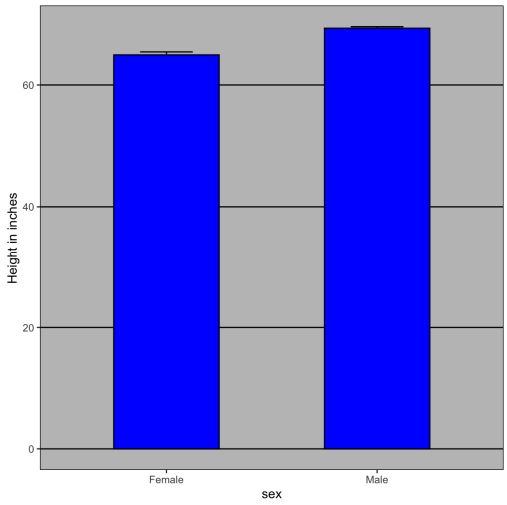
Before Starting

- Who for?
  - Need to know your audience
    - Examples: quantitative, non quantitative, other economists, your boss, grandma
- What for?
  - Need to know the objective
    - Examples: summarize the data, compare different variables or data sets, describe relationships, and so on

Show the Data

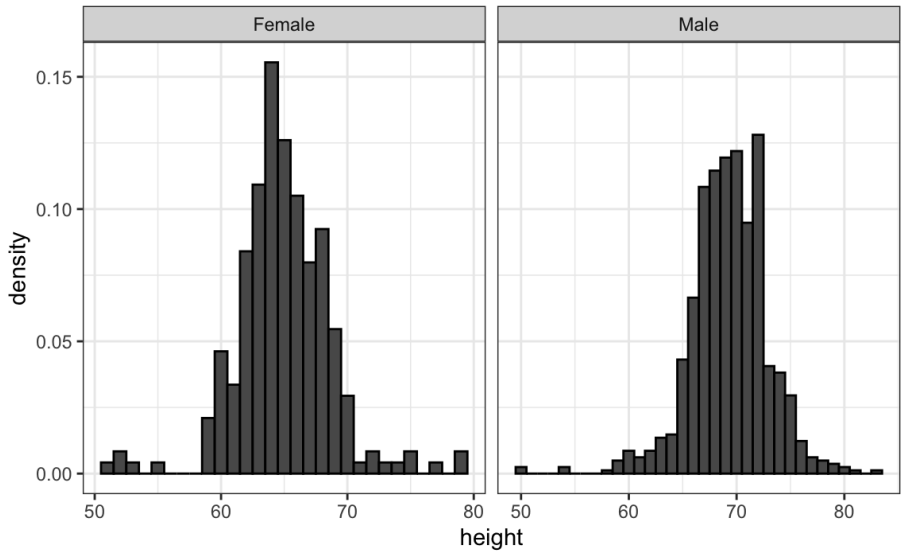


Original Data



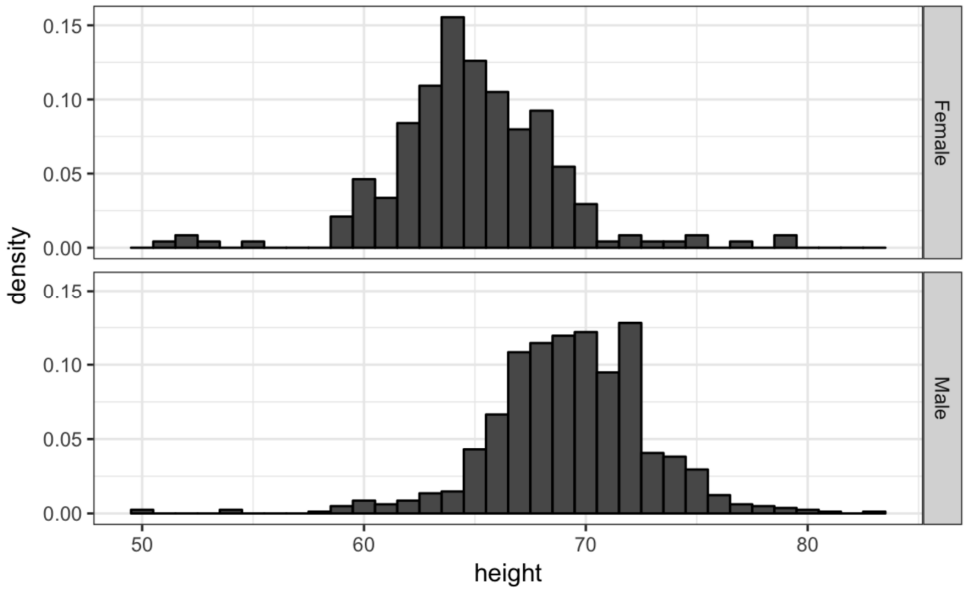
A Bar Chart With Averages Is not Very Informative

Show the Data

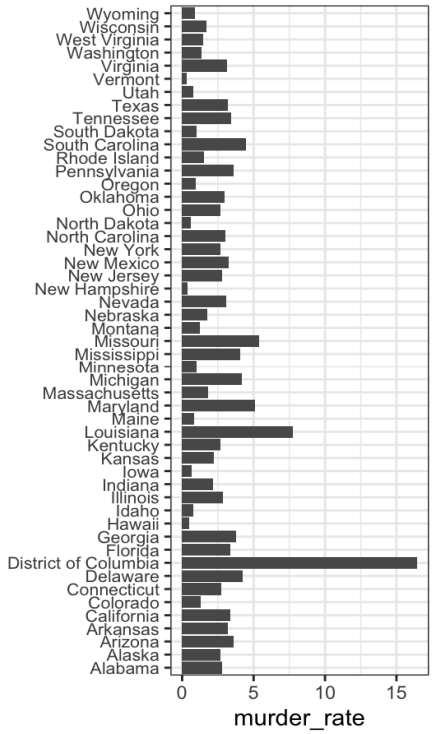


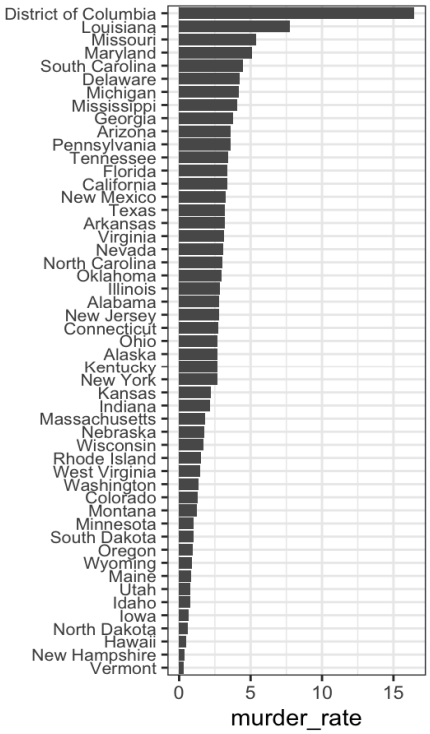
Showing the entire distribution is even better

Ease Comparisons: Align Plots

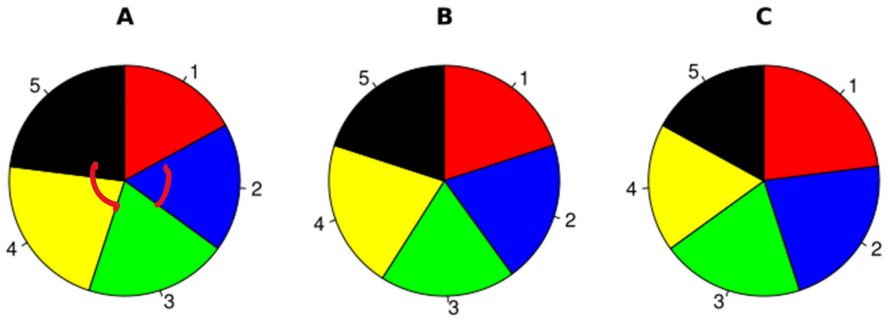


Ease Comparisons: Sort Logically Not Alphabetically



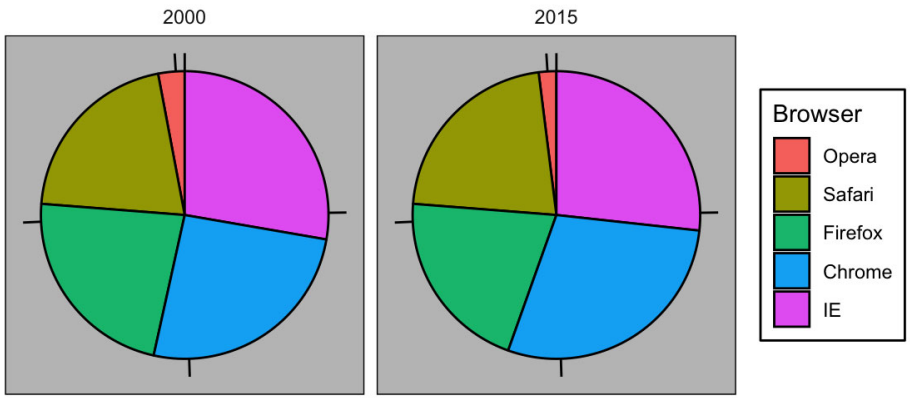


Pie Charts Are Not Good



Do you see differences between these pie charts?

Pie Charts Are Very Bad



Compare browser usage in two years: only 10 numbers! Can you see the differences?

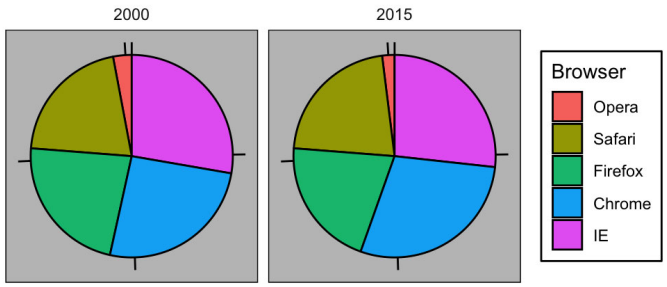
The human brain has a hard time with angles, particularly when areas are the only visual clue

Just Print the Numbers

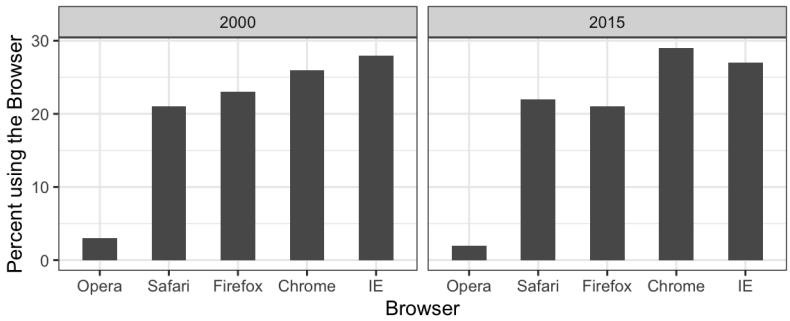
Browser	2000	2015
Opera	3	2
Safari	21	22
Firefox	23	21
Chrome	26	29
Explorer	28	27

Compare browser usage in two years: only 10 numbers! Can you see the differences?

Not All Charts Are Equal

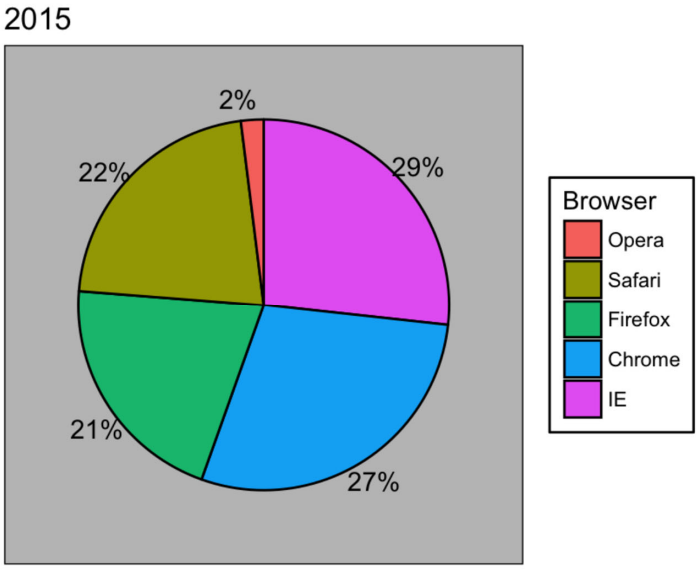


Compare browser usage in two years: only 10 numbers! Can you see the differences?



Bar Charts do a better job in this case

If You Have To



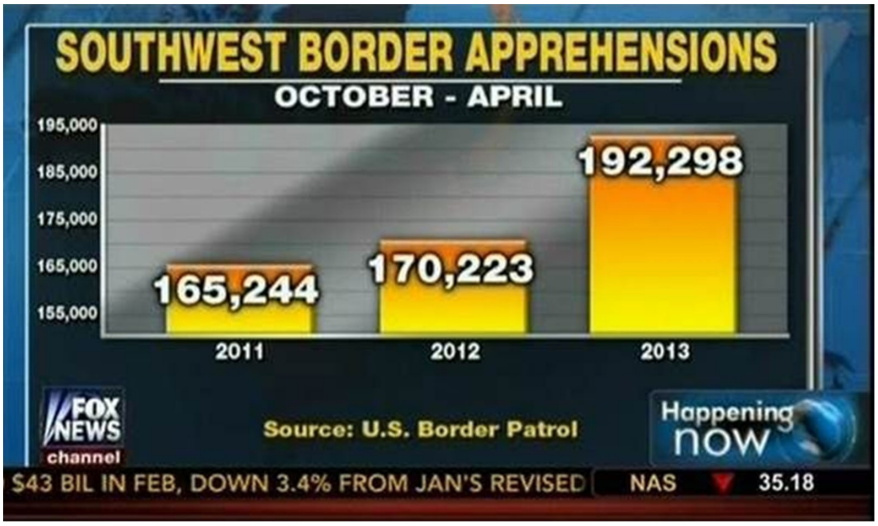
Use labels for each of the values to help the viewer understand the data better

Bad Charts Examples

- Next we will see some bad charts and try to understand what makes them bad

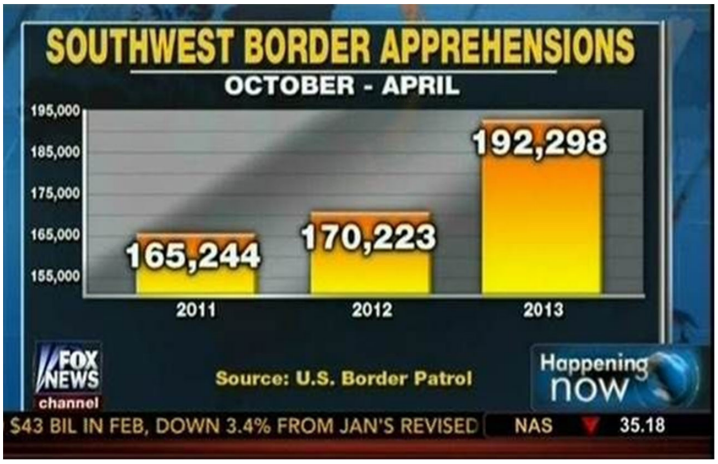
BAD = does not describe data accurately

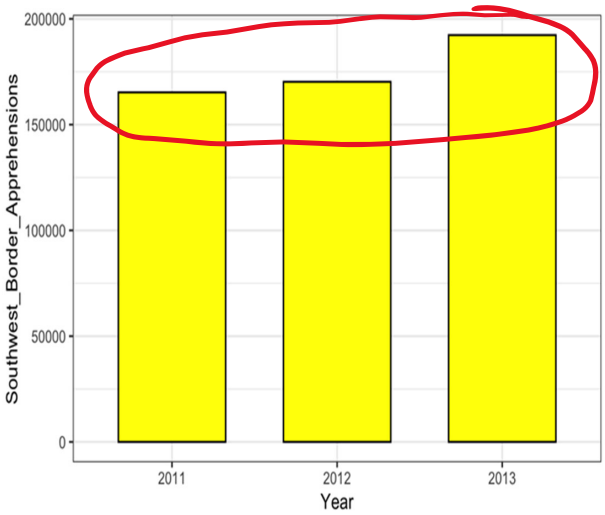
Why Is This Graph Bad?



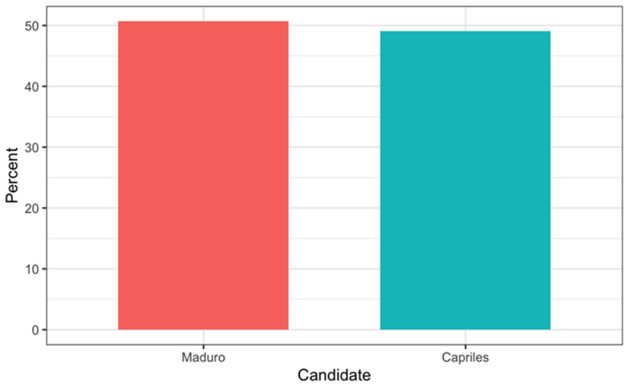
A. Because it is on tv
B. Because it uses too much color
C. Because it does not include the origin
D. Because the data is unreliable

Showing the Origin Describes the Data Better





Include the Origin

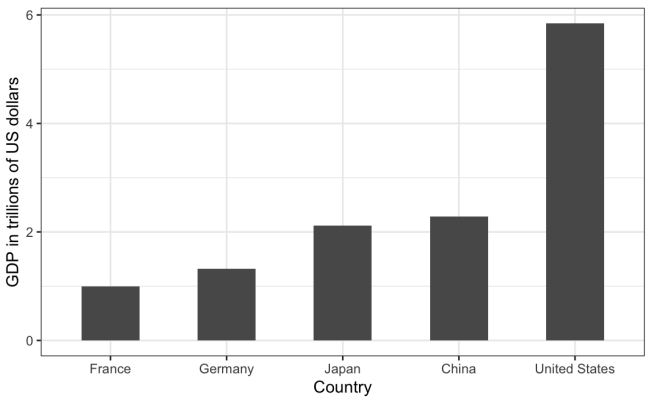


The difference in vote share is very small

Why Is the Graph On the Right Bad?

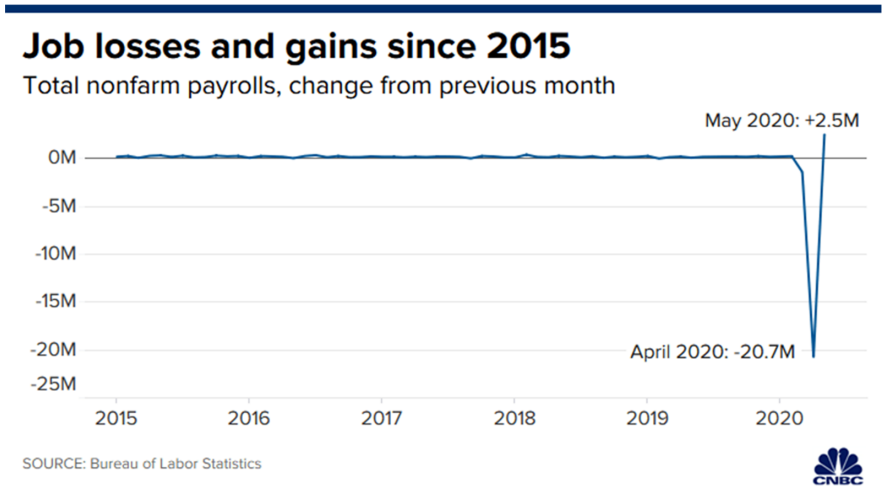
A. Because it uses circles
B. Because it uses too little color
C. Because it has a politician on the left
D. Because it distorts quantities

Do Not Distort Quantities



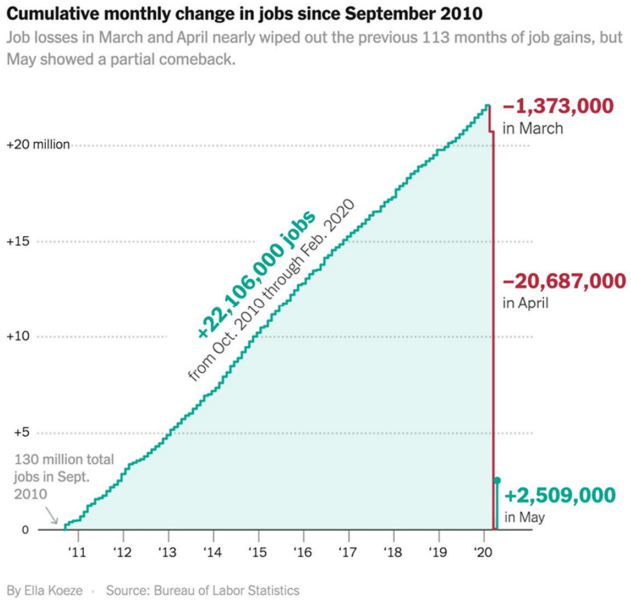
The size of the circles does not represent the actual numbers

Why Is This Graph Bad?



A. Because it implies jobs are back to normal in May 2020
B. Because the data is unreliable
C. Because it includes the origin
D. Because if excludes farms

Do Not Mislead With Data



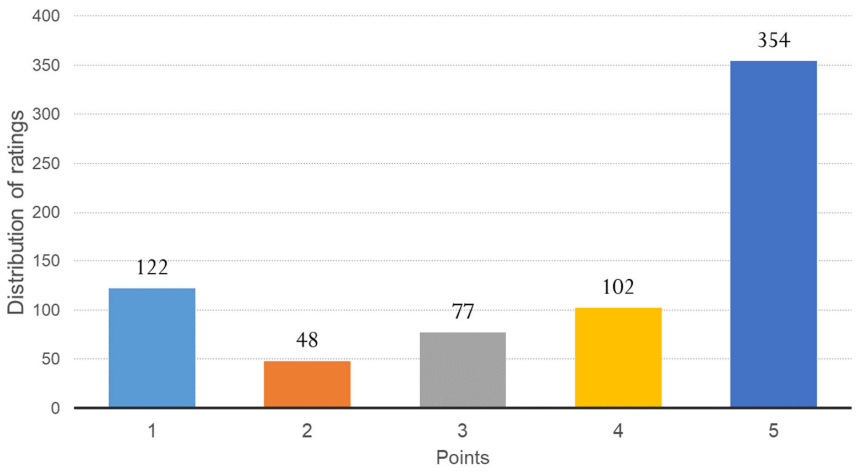
An accurate description of the job situation shows that things in May 2020 are much worse than they were in February of that same year

Histogram

- A histogram is a commonly used representation of the distribution of numerical data
- To construct a histogram, divide the entire range of values into a series of intervals (called "bins") and then count how many values fall into each interval
  - The bins are usually consecutive, non-overlapping, and are often of equal size
- A rectangle is erected over each bin with height proportional to the number of cases in each bin (called "frequency")

Histogram from Previous Slides: app ratings

Each bin represent a rating, and the height of each bar represents the number of times a particular rating has been given



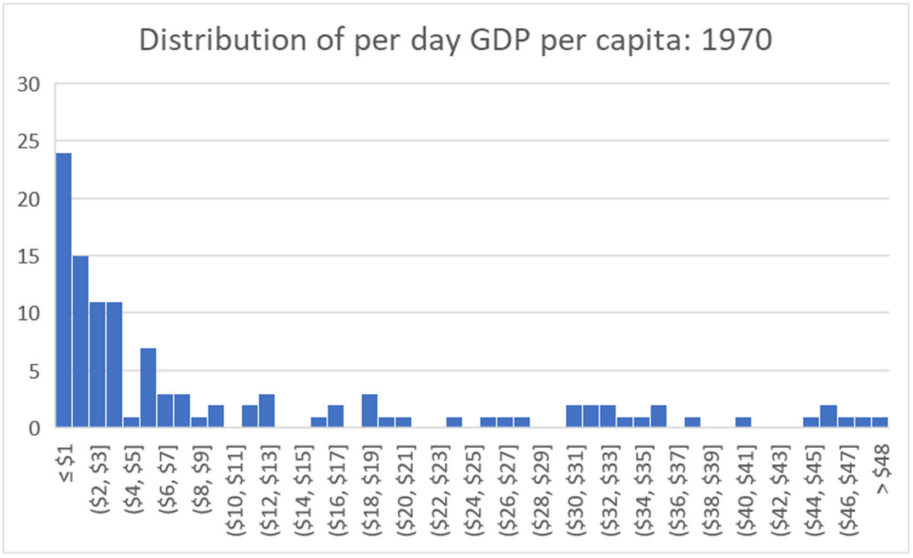
An Example: Tracking GDP Across the World

- We have data on per day Gross Domestic Product per capita in 1970 and 2010
- What can we learn about the distribution of this variable and its changes over time
- We can do this by looking at a few histograms

Income Distribution Among Countries

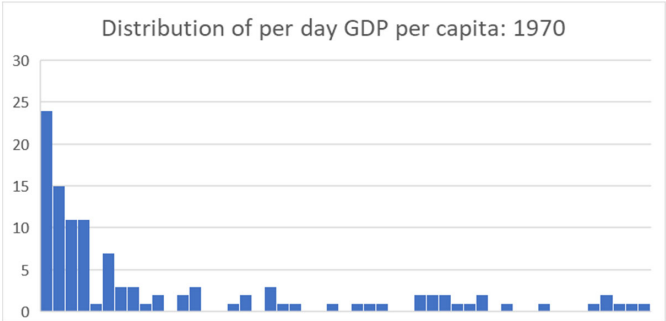
We have data on 180+ countries, and we can look at the daily per capita gross domestic product

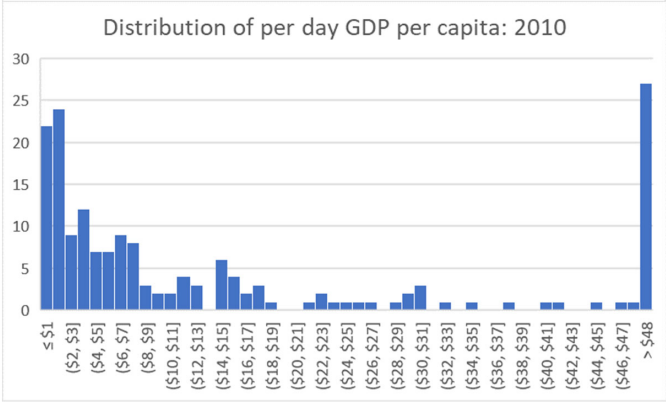
What do we see in this graph?



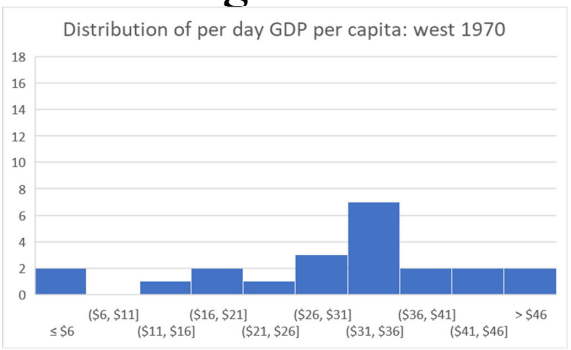
Income Distribution Among Countries Over Time

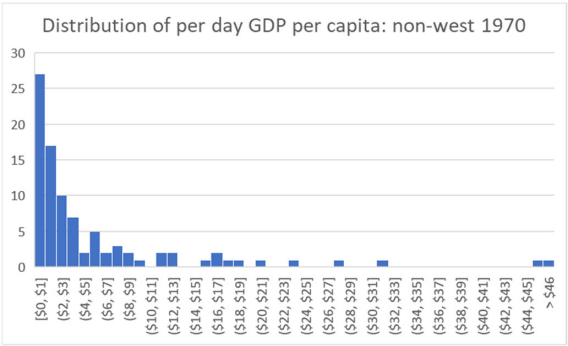
What do we see comparing the two graphs?

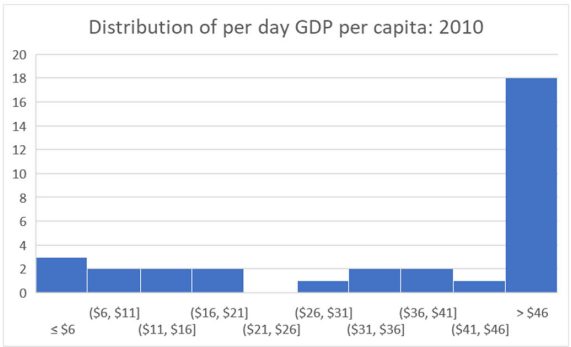


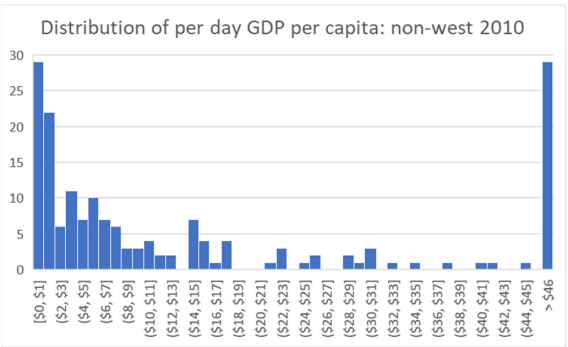


Change Over Time for West and non-West Countries

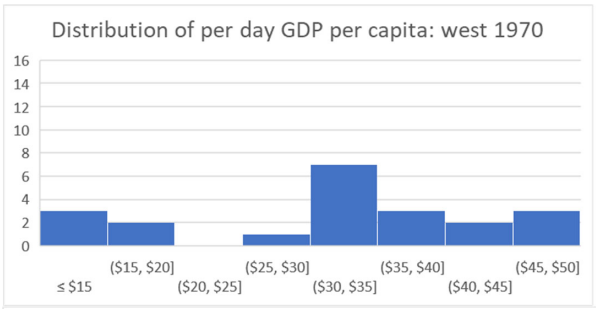


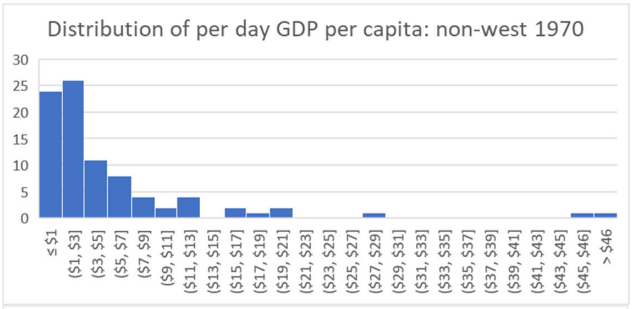


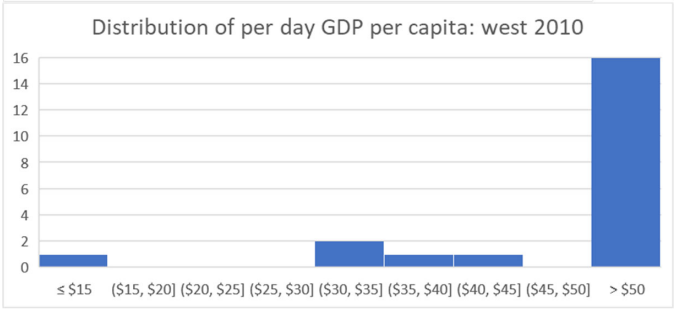


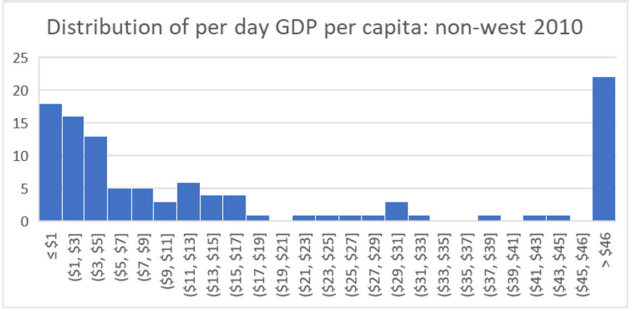


Same Analysis Excluding Incomplete Data Countries





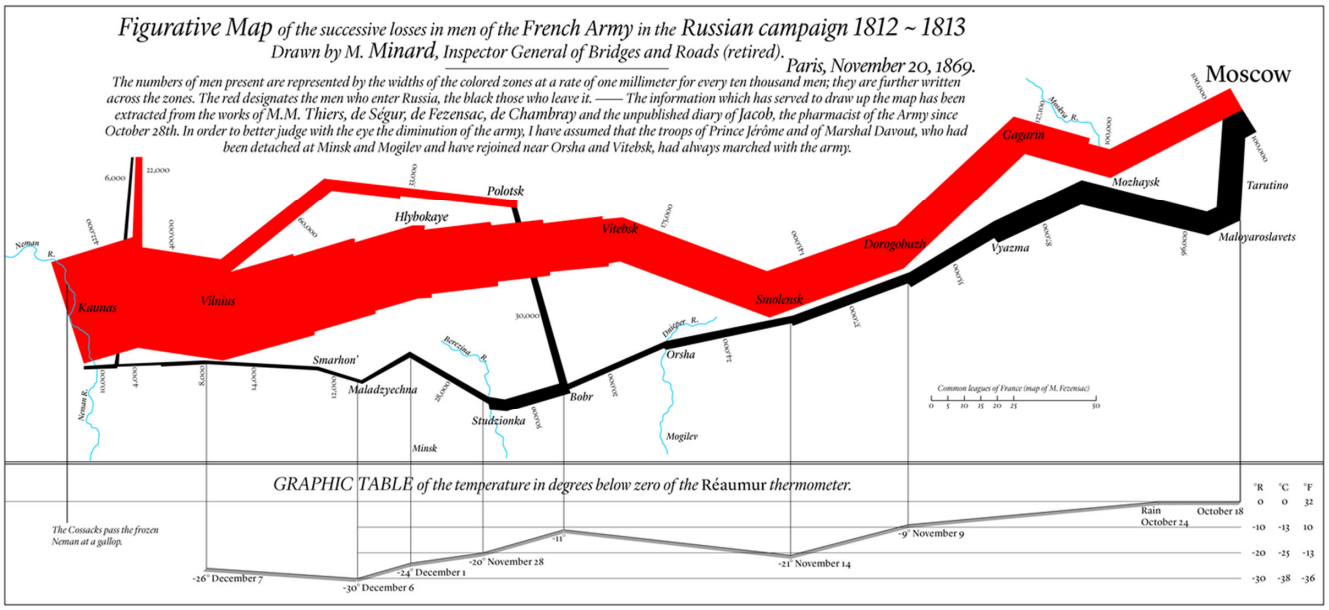




Some Principles of Graphical Excellence

- Show the data
- Induce the viewer to think about substance
- Avoid distortion
- Present many numbers in a small space
- Encourage comparisons
- Serve a purpose

The Best Graph Ever Made?



Class 2: useful facts

- Mean: *the mean of* $x_1, x_2, \dots, x_N$ *is* $\mu = \frac{x_1+x_2+\cdots+x_N}{N}$
- Variance and standard deviation:

  the variance of $x_1, x_2, \dots, x_N$ *is* $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}$

  the standard deviation of $x_1, x_2, \dots, x_N$ *is* $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}}$
- Principles of Graphical Excellence
  - Show the data
  - Induce the viewer to think about substance
  - Avoid distortion
  - Present many numbers in a small space
  - Encourage comparisons
  - Serve a purpose

Announcement

- Online Quiz 2 is due before next class at 9am

Next Time

- Introduction to Probability
  - Events
  - Union and Intersection
  - Probability Tables