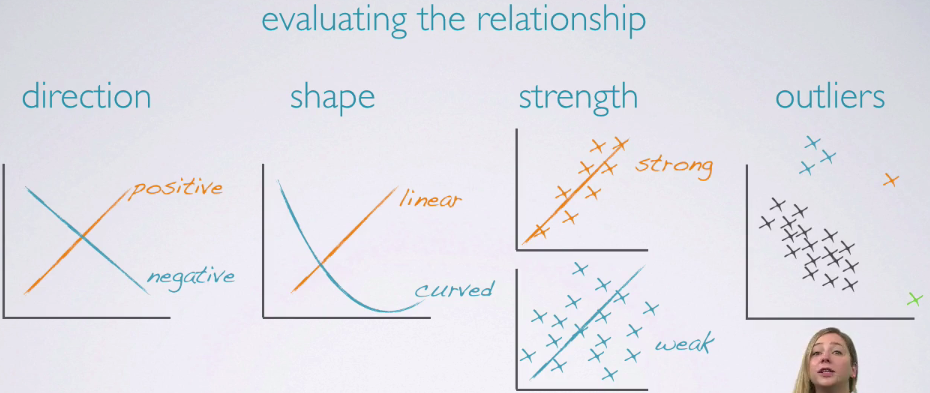
***COURSERA: STATS W/ R SPECIALIZATION***

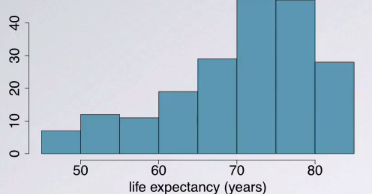
***COURSE 1 - Introduction to Probability and Data***

**WEEK 2- Exploratory Data Analysis and Introduction to Inference**

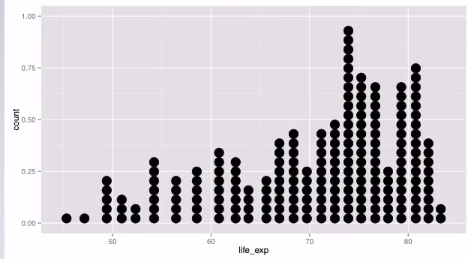
* Visualizing numerical data:
* **Scatterplots** (i.e. suspect income per person, or GDP/capita influences life expectancy, we plot life expectancy (response) as a function of GDP/capita (explanatory)
* Only correlations, not causal explanations



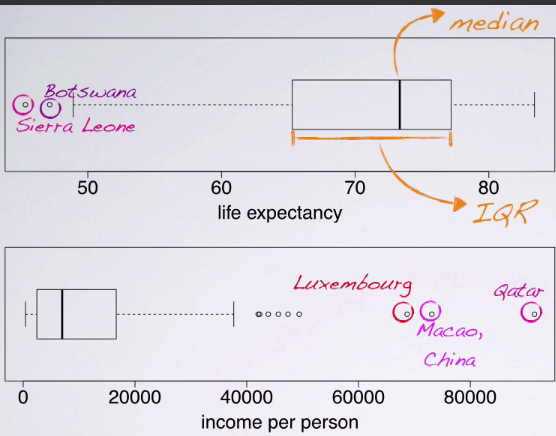
* **Outliers** may be interesting cases, so they must be handled them w/ careful consideration of the research question and of other associated variables
* **Histograms** 🡪 distributions of numerical variables = a view of **data density** (higher bar = more common data points)



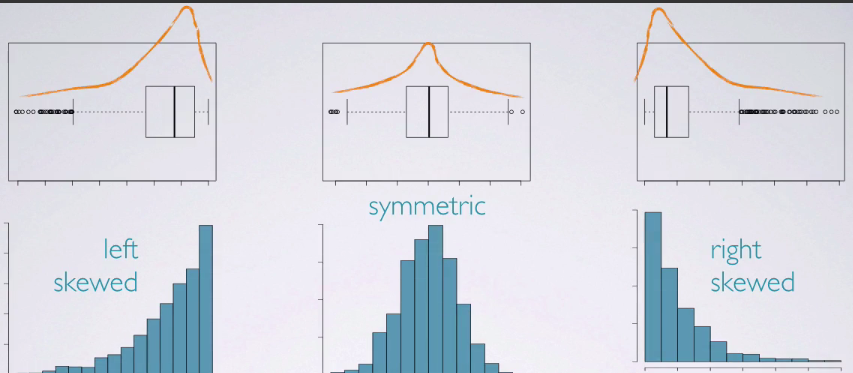
* Whether a histogram is unimodal, bimodal, multi-modal, or uniform is important
* uniform = each value has an equal chance of being represented
* Bin width/size can also alter the story a histogram tells
* Too wide = lose interesting data
* Too narrow, difficult to get overall picture of distribution
* Ideal width depends on data 🡪 play w/ it
* **Dot plot**

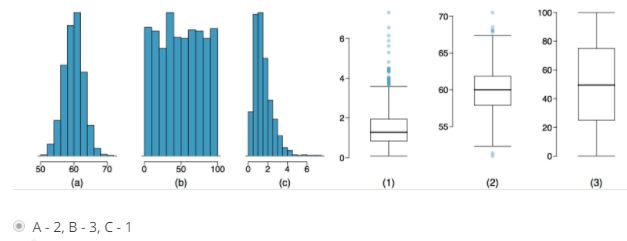


* Useful when individual data points are interesting, but can get busy w/ large sample sizes
* **Boxplots**
* Helpful to find outliers 🡪 displays median, IQR, max + min, outliers, skewness, but NOT modality (check a histogram)

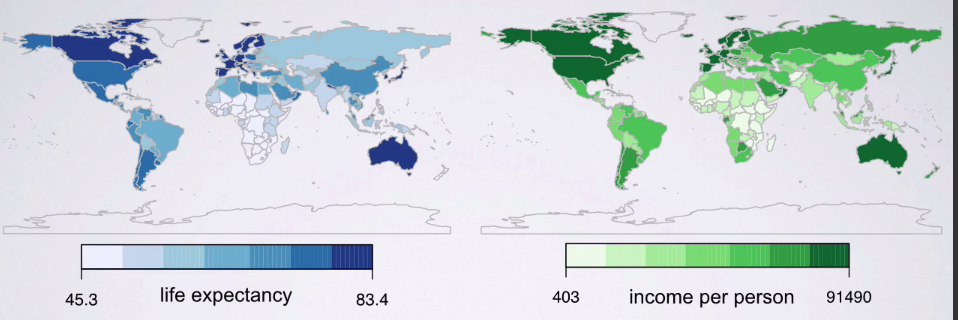


* To see skewness in boxplots, imagine its histogram 🡪 peak is around median and tails extend out to tails of boxplots





* **Intensity maps** of spacial distributions can also be helpful



Ex: both income + life exp. are lower in Africa and higher in N. America and Europe

* **Sample statistics = point estimates** of **population parameters** (which are almost always unknown b/c it’s not feasible to have info on all observations in a population)
* Sample stats might not be right, but should be good estimates *if sample is good* (i.e. representative of the population)
* Note: Latin = sample, Greek = population
* W/ continuous distributions, may be very unlikely to observe the same value multiple times, so **mode** isn’t very useful
* **Mean** is pulled to the side of the tail (below **median** in left skews, above it in right skews)

