the shortcut × moz://a

DATA SCIENCE

TRAINING PROGRAM

Week 2: Descriptive Analysis and Basic Visualization

Jesse McCrosky, Mozilla

2019-10-14

kodit.io selko SILO.AI

Descriptive Analysis: The Art & Science

- Descriptive Analysis can become quite intuitive (“artistic”) - there are so many things to look at that intuition can guide a search.
  - One finding in the descriptives might suggest another place to look
- However, to start out a more “scientific” exhaustive approach is very useful and educational.
  - Look at everything!

Variables

- Remember, variables are the “columns” in your data table
- Types of descriptive metrics and tools:
  - Univariate – “one variable”
    - How is the distribution of one particular variable across your data?
  - Bivariate – “two variables”
    - What is the relationship between two particular variables in your data?
  - Multivariate – “multiple (typically three or more) variables”
    - What is the relationship among some set of particular variables in your data?

Types of Variables (from last week)

- **Dichotomous/Binary**: is either “yes or no”, “true or false”. May be represented as words (“true” and “false”) in data or numbers (1 and 0).
- **Categorical**: takes on one of a set of possible values with no order between them. May be represented and text or integers.
- **Ordinal**: like categorical, but values have some order. Most common example is likert scale: (“strongly disagree”, “disagree”, …, “strongly agree”).
- **Interval/Ratio**: numbers in which you can measure the difference between any two values. Is “Ratio” if it is meaningful to say that one value is, for example, twice as big as another.

Different kinds of numbers: interval/ratio

- Interval versus ratio variables
  - For interval variables, only differences are meaningful; for ratio variables, ratios are meaningful as well and there is a meaningful zero
  - A date can be one day greater than another date, but not “twice as big”
  - A person can be 2 cm taller than another person, or twice as tall

Different kinds of numbers: continuous/discrete

- Discrete versus continuous variables
  - Discrete variables take on a countable set of values. Continuous variables can take any value in an interval.
  - Number of students is discrete, area of classroom is continuous
  - Really, everything is discrete once measured
  - We often “bin” continuous variables to make them discrete
    - Can also bin discrete variables! Us “age category” instead of “age in years”

Univariate Metrics for Binary Variables

- Proportion True / Proportion False
- Count True / Count False
- Variance

All of these take a set of Binary values and produce a single number!

USER_HAS_ACCOUNT
TRUE
FALSE
FALSE
TRUE
TRUE

Proportion TRUE = **0.6 or 60%**

Count FALSE = **2**

Variance = **1.2**

Univariate Metrics/Tools for Categorical Variables

- Frequency or Proportion Table
- Also potential to combine categories
  - European/Asian/American?

NATIONALITY
Canadian
Finnish
Russian
Portuguese
Finnish
Spanish

Canadian	Finnish	Russian	Portuguese	Spanish
1 (16.7%)	2 (33.3%)	1 (16.7%)	1 (16.7%)	1 (16.7%)

Univariate Metrics/Tools for Ordinal Variables

- Frequency or Proportion Table
- Proportion of responses in some set of values
- Pretending that it's a number
  - **BAD!** But sometime useful

LIKERT_DATA_IS_COOL
Strongly Agree
Neutral
Disagree
Agree
Strongly Agree
Neutral

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
0 (0%)	1 (16.7%)	2 (33.3%)	1 (16.7%)	2 (33.3%)

Proportion with any agree = 0.5 or 50%

Mean (if "Strongly Disagree" = 1, "Disagree" = 2, etc.) = 3.7 (between "Neutral" and "Agree")

Univariate Metrics/Tools for Interval/Ratio Variables

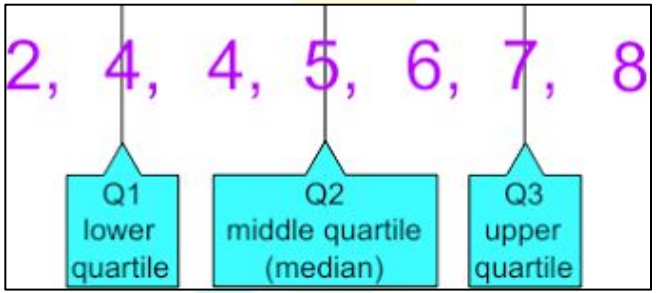
- Measure of central tendency:
  - Mean - “average” = SUM(numbers) / COUNT(numbers) = **88.3**
  - Median - “number in the middle” = **24**
  - Mode - “most common number” = **26**
- Min = -3 (problems with our inference model?)
- Max = 648 (problems with our inference mode
- Variance = 393376
- Standard Deviation = 198

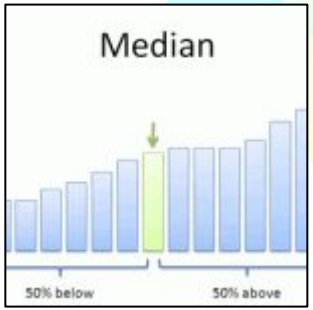
$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

INFERRED_AGE
26
8
86
15
-3
22
54
1
26
648

Quantiles

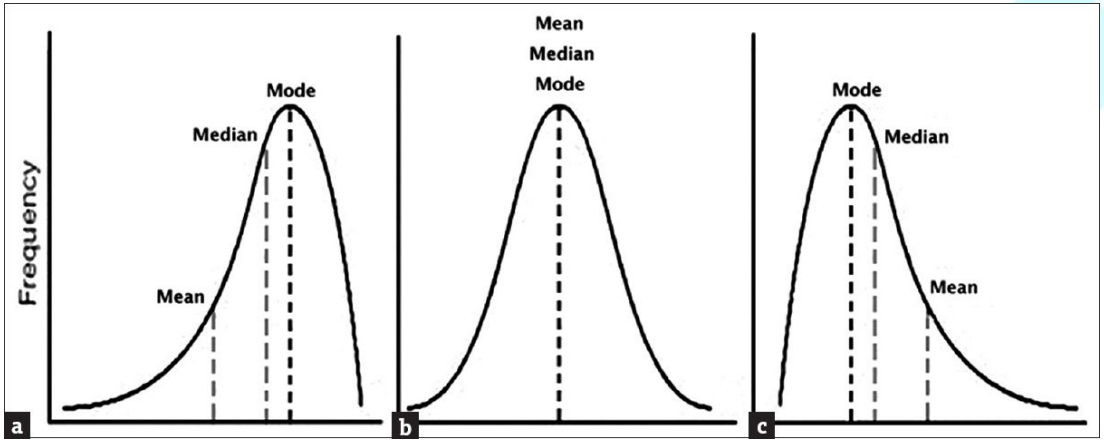
- Quantiles divide the data into intervals according to the proportion of values that fall on either side:
  - 0% quantile = minimum
  - 25% quantile = “lower quartile”
  - 50% quantile = “median”
  - 75% quantile = “upper quartile”
  - 100% quantile = “maximum”
- If a quantile falls between two values, generally, we take the mean of the two
- Advanced topic: quantiles of probability distributions





Measures of central tendency and skew

- Mean is not “robust” to “skewed distributions”
- Median is often used in these cases
- The relationship between mean and median tells you something about your data



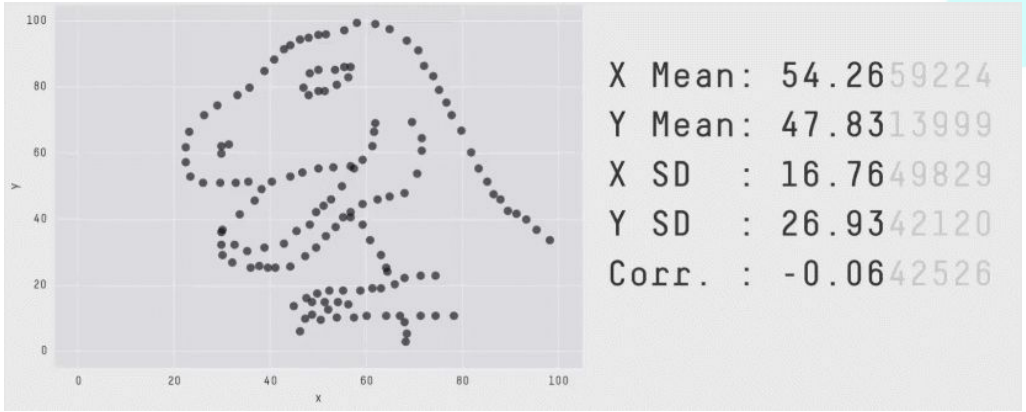
Bivariate Metrics

Examine relationship between two variables.

- Correlation
- Sliced Univariate Metrics
- Crosstabs

Correlation

- Pearson Correlation Coefficient: Most commonly used - evaluates linear relationship only
- Spearman Rank Coefficient: More robust to nonlinear relationships - evaluates rank order
- But...



Correlation and Causation

- Why do we care about correlation?
  - Suggests some important relationship between variables
  - e.g. App Store review Stars and probability of user retention
- But correlation in not causation!
  - Nurses study and hormone replacement therapy example
  - Does a high-performance computer cause people to use the internet more?

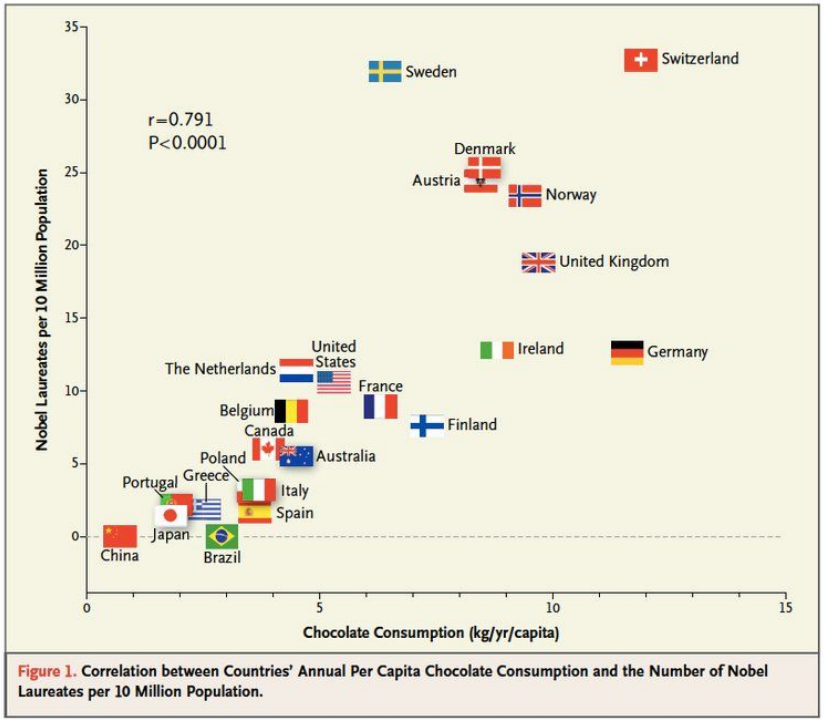
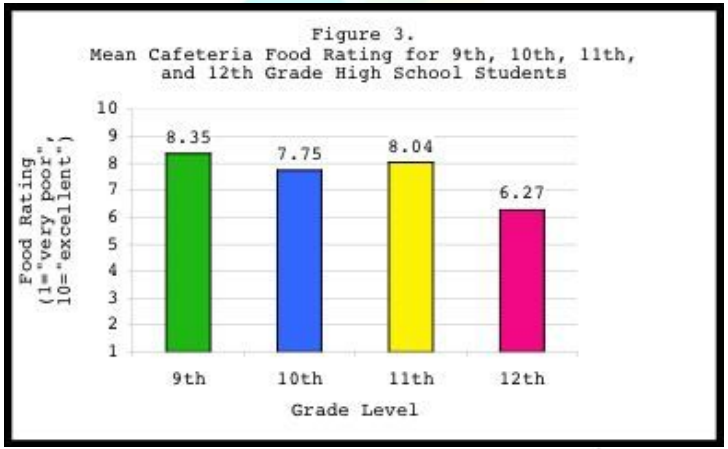


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Slicing

- We can also calculate univariate statistics, conditioned on another variable
- Referred to as “slicing” or “segmentation”
- The variable you slice by can be called a “dimension”
- Useful for finding relationships between variables or also distinct sub-populations in your data
    - Maybe data from Finnish and Canadian users are quite different

Grade	9	10	11	12
Mean Rating	8.35	7.75	8.04	6.27
Standard Deviation	4.3	4.1	3.1	2.8



Crosstabs

- A table in two dimensions can describe two variables
- Each cell is a count of how many observations (rows) have the particular values for the two variables included
- Useful for seeing general distribution of data as well as look for relationships

Class rank * Do you live on campus? Crosstabulation

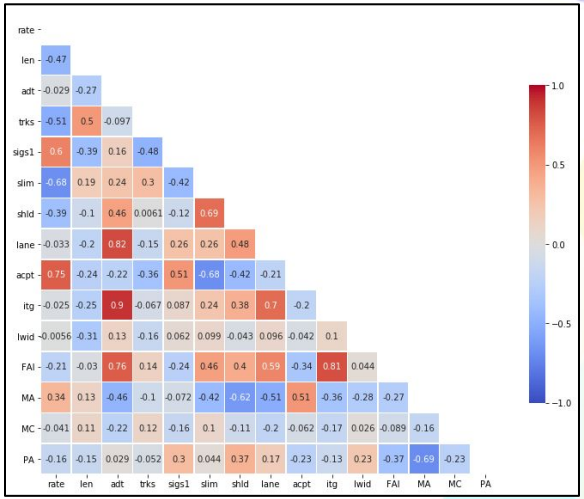
Count

		Do you live on campus?		
		Off-campus	On-campus	Total
Class rank	Freshman	37	100	137
	Sophomore	42	48	90
	Junior	90	8	98
	Senior	62	1	63
Total		231	157	388

Cross tabulation Frequency Percent		What is Your Favorite Baseball Team?			
		Toronto Blue Jays	Boston Red Socks	New York Yankees	Row Totals
In What City Do You Reside?	Boston, MA	11	33	7	51
	Row Percent	21.57%	64.71%	13.73%	34.93%
	Montreal, Canada	23	14	9	46
	Row Percent	50.00%	30.43%	19.57%	31.51%
	Montpellier, VT	22	13	14	49
	Row Percent	44.90%	26.53%	28.57%	33.56%
	Column totals	56	60	30	146
	Column Percent	38.36%	41.10%	20.55%	100.00%

Multivariate Metrics

- Can do tables of three or more variables
- Correlation Matrices can be useful to get an overview of the relationships in the data
- Principal Component Analysis (Factor Analysis) is a method to reduce the number of variables in the data - advanced topic



Period	Island		Causes		Total
			Known	Unspecifier	
Inter-fire period (1 May - 31 October)	Skopelos	Count	22	5	27
		Expected Count	17.1	9.9	27.0
	Skiathos	Count	5	14	19
		Expected Count	12.0	7.0	19.0
	Alonnisos	Count	30	14	44
		Expected Count	27.9	16.1	44.0
	Total	Count	57	33	90
		Expected Count	57.0	33.0	90.0
No-fire period (1 November - 30 April)	Skopelos	Count	21	3	24
		Expected Count	19.2	4.6	24.0
	Skiathos	Count	6	6	12
		Expected Count	9.6	2.4	12.0
	Alonnisos	Count	25	4	29
		Expected Count	23.2	5.8	29.0
	Total	Count	52	13	65
		Expected Count	52.0	13.0	65.0

The Art of Descriptive Analysis

- Feel and intuition
  - Use domain knowledge
  - Look for patterns
  - Keep notes
  - Look for outliers and potential data issues

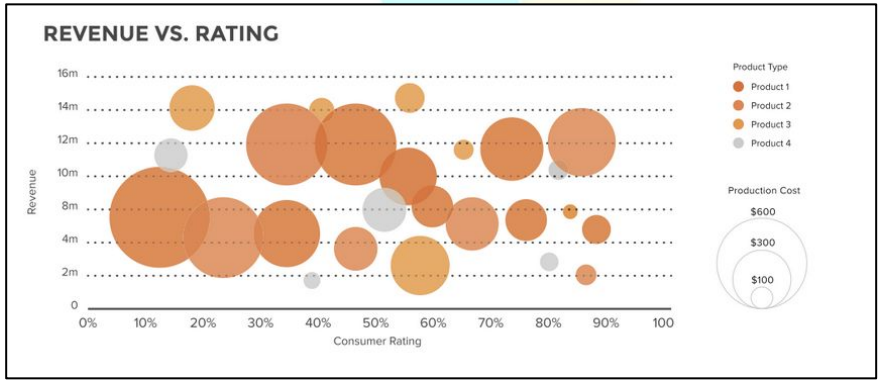
BREAK

Basic Visualization

- A picture is worth 1000 numbers?
  - Visualization can express data very effectively, but can also be very confusing (or worse, misleading) if done poorly
  - Always think of what you're trying to express
  - Every element of the visualization should be chosen for a reason
  - Dimensions: x-axis, y-axis, color, size, etc.?

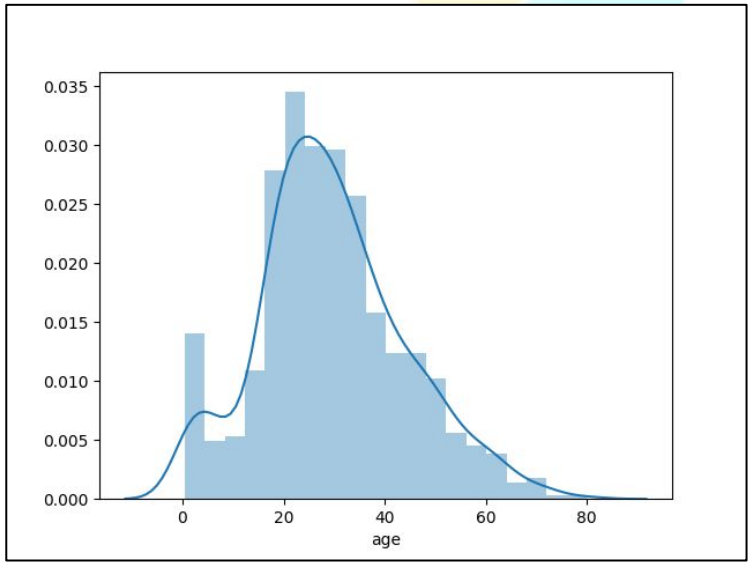
Types of basic visualization:

- Univariate distribution plots (histograms, densities)
- Scatter plots
- Summary statistic plots (bar charts)



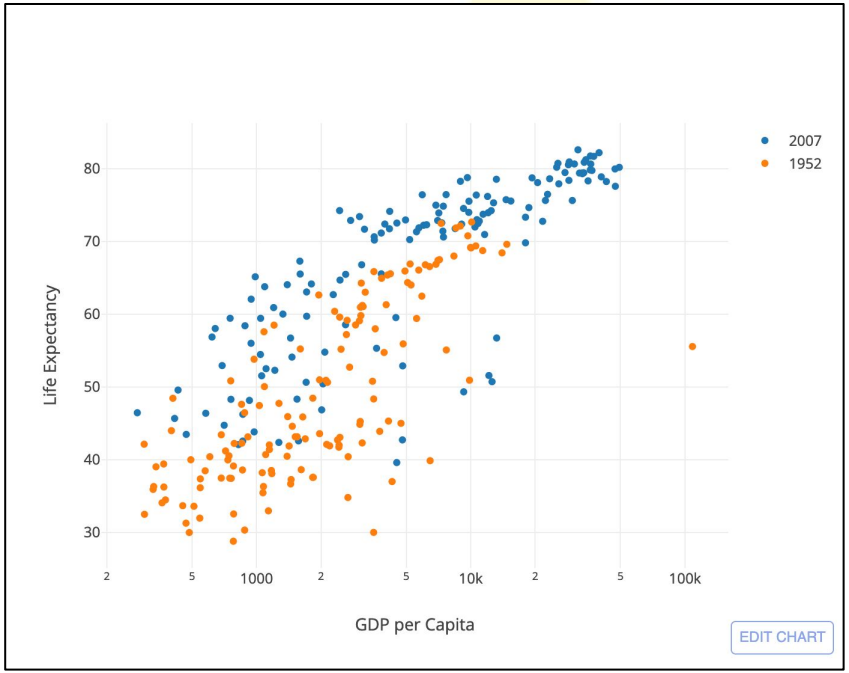
Univariate Distribution Plots

- Useful for discrete variables (or binned continuous variables)
- X-axis is the variable we want to see the distribution of
- Y-axis is density “how often variable takes that value” (ax axis should be labeled?)
- Bars represent exact counts from data
- Line represents smoothed “KDE” estimate of distribution



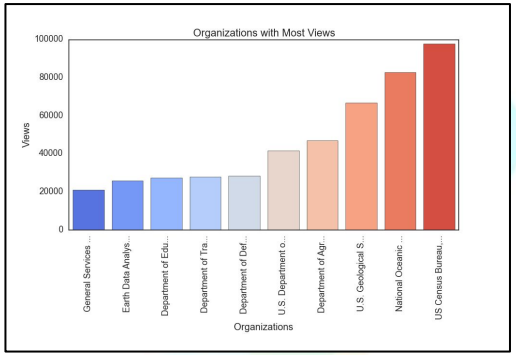
Scatter Plots

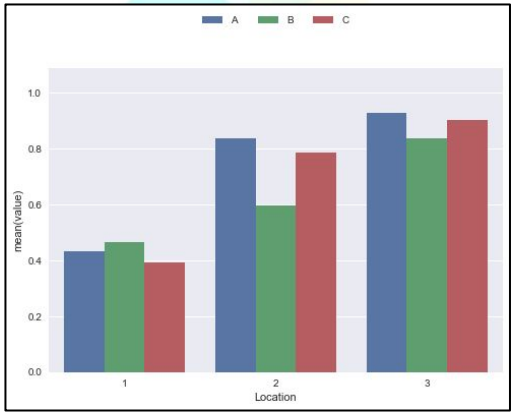
- Shows one marker for each row of data with variables for x-axis, y-axis, and sometimes color (as in this case), marker style, or size.
- Good overview of distribution of 2+ variables as well as relationship
- Note log scale on x-axis



Summary Statistic Plots

- Shows summary statistic (remember univariate metrics?) on one axis by some discrete variable on another
- Complexities like multi-bars, stacked bars, etc.





Preparing for the lab

Open this notebook.

Onwards

- Learn this in the lab!
- Questions?