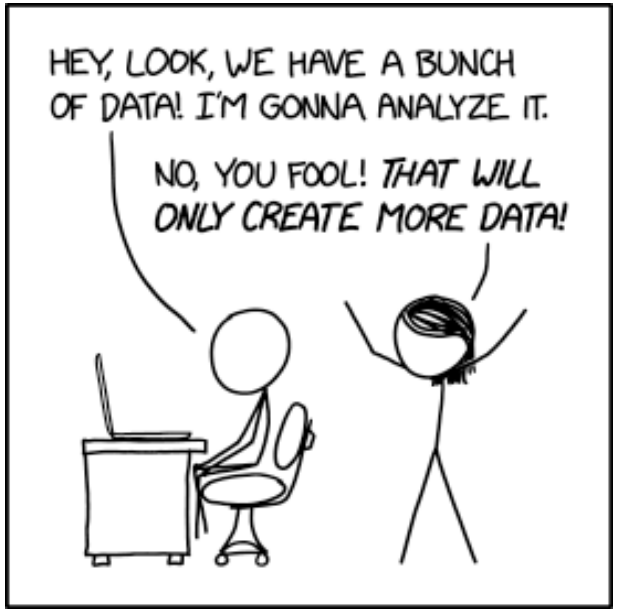
Exploratory data analysis (EDA)

Introduction to data analysis: Lecture 3

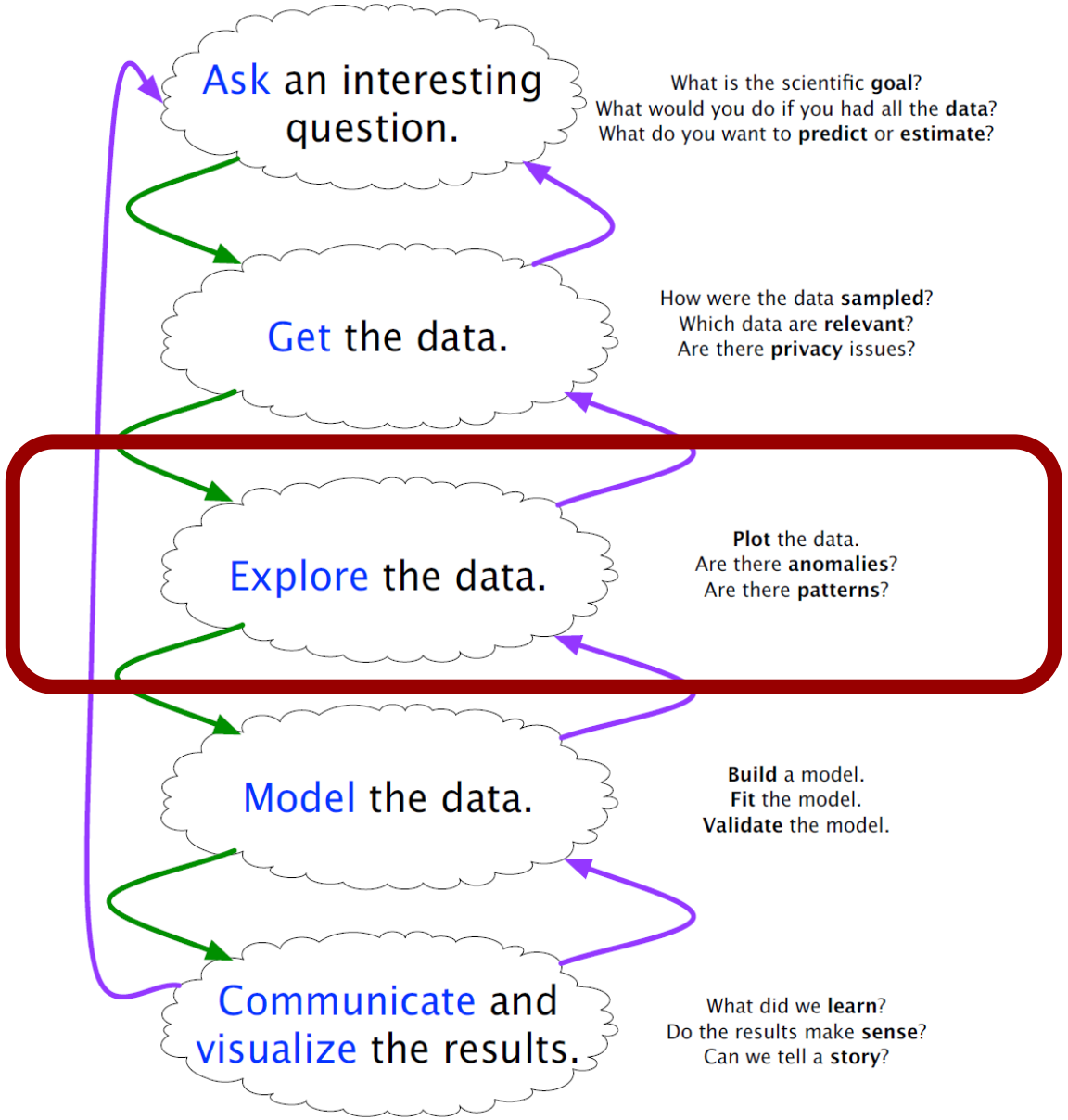
Ori Plonsky

Spring 2023



Source: xkcd

Slides partially based on Harvard CS109, Berkeley data8, and data-in-a-box


Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

This lecture

- Population and sample
- Tidy (and untidy) data
- Representing data
- Descriptive statistics
- Feature engineering

Population vs. Sample

Data often includes a sample from a larger population

- Population includes **all** the elements (individuals) of interest
- Sample is a **subset** of observations from a population

Working with data

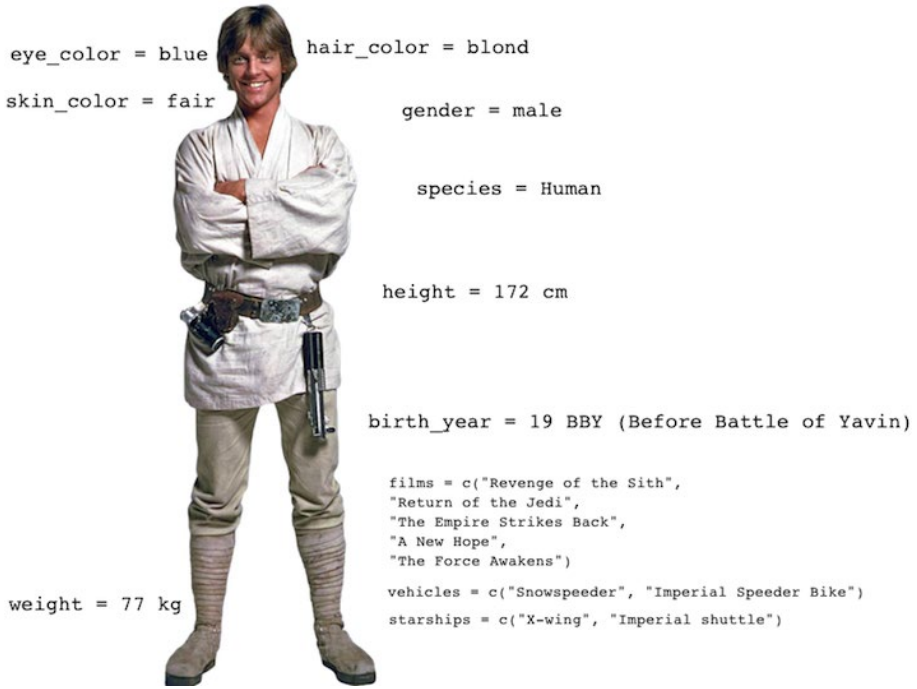
We will generally use tables to store our data:

- Each row is an observation
- Each column is a feature or variable

Working with data

We will generally use tables to store our data:

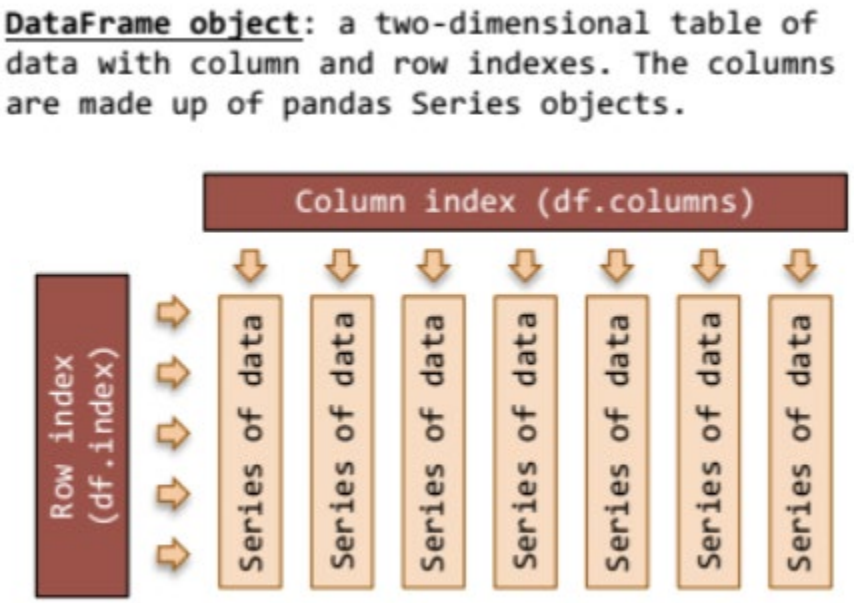
- Each row is an observation
- Each column is a feature or variable



name	height	Weight	hair_color	skin_color	eye_color	birth_year	gender	homeworld	species
Luke Skywalker	172	77	blond	fair	blue	19.0	male	Tatooine	Human
C-3PO	167	75	*NA*	gold	yellow	112.0	*NA*	Tatooine	Droid

Tidy data

- We expect each row to represent a single object or event
  - Survey responses; movie characters; transaction; Twitter accounts
- We expect each column to represent a single attribute of the object
- We expect each table to contain only observations of the same kind



Un-tidy data

- Messy format
  - Column headers are values, not variables (wide vs. long format)
  - Multiple variables in one column
  - Multiple observational units in the same table
- Wrong values
- Missing values
  - Ignoring?
  - Removing incomplete observations?
  - Filling in?

- Tidying up the data is part of EDA

Types of variables

- Numeric

- Boolean (True-False / binary)

- Categorical

Types of variables

- Numeric
  - continuous (float) – infinite number of values in any bounded interval
    - Height
  - discrete (integer) – finite number of value in any bounded interval
    - Number of siblings
- Boolean (True-False / binary)
  - Can be (and often is) treated as numeric (0/1)
- Categorical
  - Ordinal - levels have a natural ordering.
    - Educational level
  - Nominal – no natural ordering
    - Which pet do you have?

Variable coding

- Often, during EDA, we code categorical variables as numeric for more convenient workflow
- This does **not** make the categorical variable a numeric one!
- This makes sense, and is often required
- But can lead to catastrophic results if not done carefully
  - Check and re-check!

Comment & Response

October 8, 2019

Notice of Retraction. Aboumatar et al. Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease: A Randomized Clinical Trial. *JAMA*. 2018;320(22):2335-2343.

Hanan Aboumatar, MD, MPH[1,2,3,4,5]; Robert A. Wise, MD[6]

Author Affiliations | Article Information

JAMA. 2019;322(14):1417-1418. doi:10.1001/jama.2019.11954

To the Editor On behalf of our coauthors, we write to report a programming error and other errors that affected the results in our article, "Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease: A Randomized Clinical Trial" published in the December 11, 2018, issue of *JAMA*.[1] We write to explain what happened and to request retraction of this article.

The identified programming error was in a file used for preparation of the analytic data sets for statistical analysis and occurred while the variable referring to the study "arm" (ie, group) assignment was recoded. The purpose of the recoding was to change the randomization assignment variable format of "1, 2" to a binary format of "0, 1." However, the assignment was made incorrectly and resulted in a reversed coding of the study groups. Even though the data analyst created and conducted some test analysis programs, they were of the type that did not show any labeling of the arm categories, only the "arm" variable in a regression, for example. After detecting this error, we promptly reported it to our institutional review board and appropriate offices within our university, alerted *JAMA*, and proceeded to confirm whether the error had affected the analytic data sets, which we found to be the case. We therefore started a complete data reanalysis, with 2 biostatisticians performing double programming and an independent analysis of study primary outcomes to ensure the validity of the reported results. As noted here, this reanalysis showed reversed study findings, with a higher number of hospitalizations and emergency department visits in the intervention compared with the usual care group.

Representing variables

Example: company choice

You got offers to join three different companies - A, B, or C - as a data scientist. To make your decision, you gather data about the salaries of data scientists in the three companies, and find the following:

Company A	15500	23500	18000	19000	21500
Company B	19500	20500	19500	19500	18500
Company C	14500	15000	15000	15000	38000

Which company will you choose? Why?

Example: company choice

You got offers to join three different companies - A, B, or C - as a data scientist. To make your decision, you gather data about the salaries of data scientists in the three companies, and find the following:

Company A	15500	23500	18000	19000	21500
Company B	19500	20500	19500	19500	18500
Company C	14500	15000	15000	15000	38000

Which company will you choose? Why?

Notebook demo:

- Save `Lecture 3.ipynb` and `salaries.csv` in the same folder
- Open `Lecture 3.ipynb` in Jupyter Notebook

Measures of centrality

- Mean: the average data value $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Median: the “middle” data value. Half of the observations are smaller, and half are larger, than the median.

- Mode: the value that occurs most often in the data
  - Often not very useful

Measures of centrality

- Mean: the average data value $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Median: the “middle” data value. Half of the observations are smaller, and half are larger, than the median.

- Mode: the value that occurs most often in the data
  - Often not very useful

- Which of these are sensitive to extreme values?

Measures of centrality

- Mean: the average data value $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Median: the “middle” data value. Half of the observations are smaller, and half are larger, than the median.

- Mode: the value that occurs most often in the data
  - Often not very useful

- Which of these are sensitive to extreme values?
  - Outliers?

Measures of centrality

- Mean: the average data value $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Median: the “middle” data value. Half of the observations are smaller, and half are larger, than the median.

- Mode: the value that occurs most often in the data
  - Often not very useful

- Which of these are sensitive to extreme values?
  - Outliers?
- Which of these are relevant for categorical variables?

Question

What is one advantage of the median vs. the mean?

A. The median is always greater than the mean
B. The median is less affected by outliers than the mean
C. The mean and median are the same thing.
D. The median is less mean than the mean

Question

It is known that, for some data, the mean of X is smaller than the mean of Y. Which of the following is **always** correct?

A. The median of X is smaller than the median of Y

B. The mode of X is smaller than the mode of Y

C. The median of X is smaller than the mean of Y

D. None of the statements above is always correct

Measures of spread

- Spread = by how much do measures of centrality describe the data
- Range = Maximum value – Minimum value
- Variance (in a sample): measures how much the values (in the sample) deviate from the mean (of the sample)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Standard deviation: the square root of the variance

Measures of spread

- Spread = by how much do measures of centrality describe the data
- Range = Maximum value – Minimum value
- Variance (in a sample): measures how much the values (in the sample) deviate from the mean (of the sample)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  - Units?
  - Sensitivity to extreme values?
- Standard deviation: the square root of the variance

Measures of spread

- Spread = by how much do measures of centrality describe the data
- Range = Maximum value – Minimum value
- Variance (in a sample): measures how much the values (in the sample) deviate from the mean (of the sample)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  - Units?
  - Sensitivity to extreme values?
- Standard deviation: the square root of the variance
  - Units?

Percentiles

- The *p*th percentile is the value below which *p*% of the observations lie
  - The median is the 50th percentile

Percentiles

- The *p*th percentile is the value below which *p*% of the observations lie
  - The median is the 50th percentile
- Q1 = the first quartile = the 25th percentile
- Q3 = the third quartile = the 75th percentile
- IQR = Inter-quartile range = Q3-Q1

Percentiles

- The *p*th percentile is the value below which *p*% of the observations lie
  - The median is the 50th percentile
- Q1 = the first quartile = the 25th percentile
- Q3 = the third quartile = the 75th percentile
- IQR = Inter-quartile range = Q3-Q1
  - Sensitivity to extreme values?

Distributions

- (Empirical) Distribution: for each possible value of a variable, how many observations have that value
  - Or, what is the proportion of observations that have that value

Linear correlation

- Pearson correlation: measures *linear* relationship between two variables
  - Measures if two variables change at a constant rate with respect to each other
  - Between -1 (perfect negative correlation) and 1 (perfect positive correlation)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

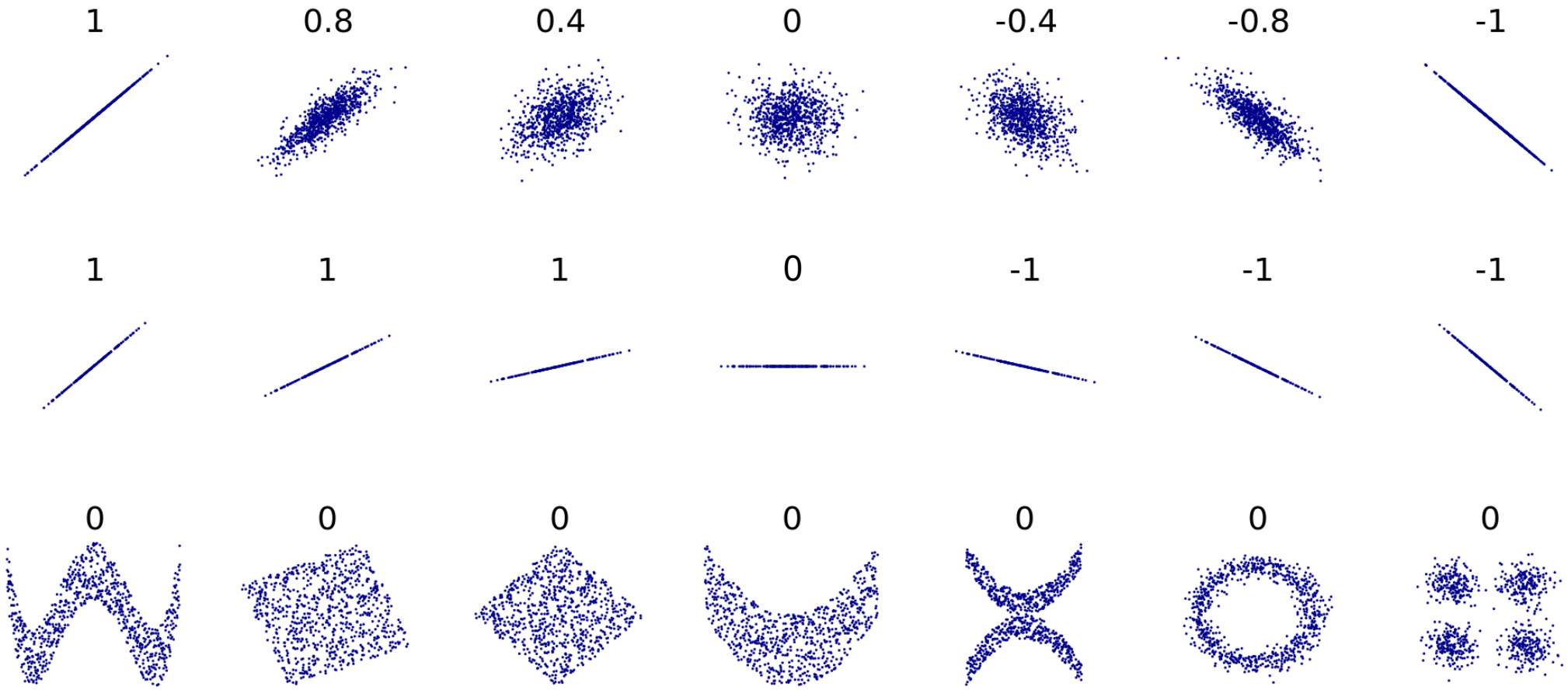
Linear correlation

- Pearson correlation: measures *linear* relationship between two variables
  - Measures if two variables change at a constant rate with respect to each other
  - Between -1 (perfect negative correlation) and 1 (perfect positive correlation)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Distinguish from *any* correlation
  - Lecture 2: “Two variables are correlated when knowing the value of one gives you information about the likely value of the other”

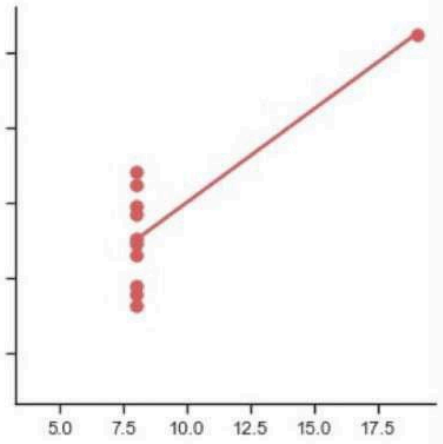
Visualizing linear correlation



Summary statistics are not enough

YES, BUT

r = 0.82



Summary statistics are not enough

- Notebook demo:
  - Save `Datasaurus.csv` in the same folder you saved `Lecture 3.ipynb`
  - Open `Lecture 3.ipynb` in Jupyter Notebook again

Re: Un-tidy data

- Messy format
  - Column headers are values, not variables (wide vs. long format)
  - Multiple variables in one column
  - Multiple observational units in the same table
- Wrong values
- Missing values
  - Ignoring?
  - Removing incomplete observations?
  - Filling in?

- Tidying up the data is part of EDA

Missing data

- Dropping rows
- Dropping columns
- Imputation = filling in missing values
  - Default?
  - With the mean? Median?
  - With the mode?
  - With category “missing”

Feature engineering

- The process of transforming raw, unprocessed data into features that best represent the underlying problem
  - and that make machine learning algorithm work
- Usually involves domain knowledge
- The goal is to get the most out the available data
- In practice, iterative process

Feature engineering

- The process of transforming raw, unprocessed data into features that best represent the underlying problem
  - and that make machine learning algorithm work
- Usually involves domain knowledge
- The goal is to get the most out the available data
- In practice, iterative process

- “Coming up with features is difficult, time-consuming, requires expert knowledge. ‘Applied machine learning’ is basically feature engineering.”

-Andrew Ng, Stanford

Sub-processes in feature eng.

- Feature extraction – automatic construction of features
  - E.g. from image, audio, or text data

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Sub-processes in feature eng.

- Feature extraction – automatic construction of features
  - E.g. from image, audio, or text data
- Feature selection – automatically choosing subset of “best” features
  - Usually involves some sort of feature importance measure

“If you don't reveal some insights soon, I’m going to be forced to slice, dice, and drill!”

Sub-processes in feature eng.

- Feature extraction – automatic construction of features
  - E.g. from image, audio, or text data
- Feature selection – automatically choosing subset of “best” features
  - Usually involves some sort of feature importance measure
- Feature learning – automatic identification of new features
  - Deep learning methods
  - Usually impossible to understand what features were “engineered”

“If you don't reveal some insights soon, I’m going to be forced to slice, dice, and drill!”

Sub-processes in feature eng.

- Feature extraction – automatic construction of features
  - E.g. from image, audio, or text data
- Feature selection – automatically choosing subset of “best” features
  - Usually involves some sort of feature importance measure
- Feature learning – automatic identification of new features
  - Deep learning methods
  - Usually impossible to understand what features were “engineered”
- Feature construction – **manually** constructing new features
  - An artform

“If you don't reveal some insights soon, I’m going to be forced to slice, dice, and drill!”

Examples of feature construction

- Decompose date-time
  - Hour of day, part of day, day of week, part of week, month, season, year…

Examples of feature construction

- Decompose date-time
  - Hour of day, part of day, day of week, part of week, month, season, year…
- Decompose categorical variables
  - Indicator for specific category
  - Sub-divide categories

Examples of feature construction

- Decompose date-time
  - Hour of day, part of day, day of week, part of week, month, season, year…
- Decompose categorical variables
  - Indicator for specific category
  - Sub-divide categories
- Numerical variable transformations
  - Unit-change, rounding (or modulo), discretization (binning)
  - Sum, difference, product, quotient, polynomial combinations of features
  - Indicators for important thresholds or events

Examples of feature construction

- Decompose date-time
  - Hour of day, part of day, day of week, part of week, month, season, year…
- Decompose categorical variables
  - Indicator for specific category
  - Sub-divide categories
- Numerical variable transformations
  - Unit-change, rounding (or modulo), discretization (binning)
  - Sum, difference, product, quotient, polynomial combinations of features
  - Indicators for important thresholds or events
- Combine with external data sources
  - Demographic data, geo-location data

Exercise: transactions data

Imagine you have a dataset with customer transactions, including:

- Customer_id
- Product
- Price
- Timestamp of purchase

Your goal is to characterize customers.

Think of features you can create from the data that would help you do that.