BINF 5003: Data Mining, Modeling, and Biostatistics

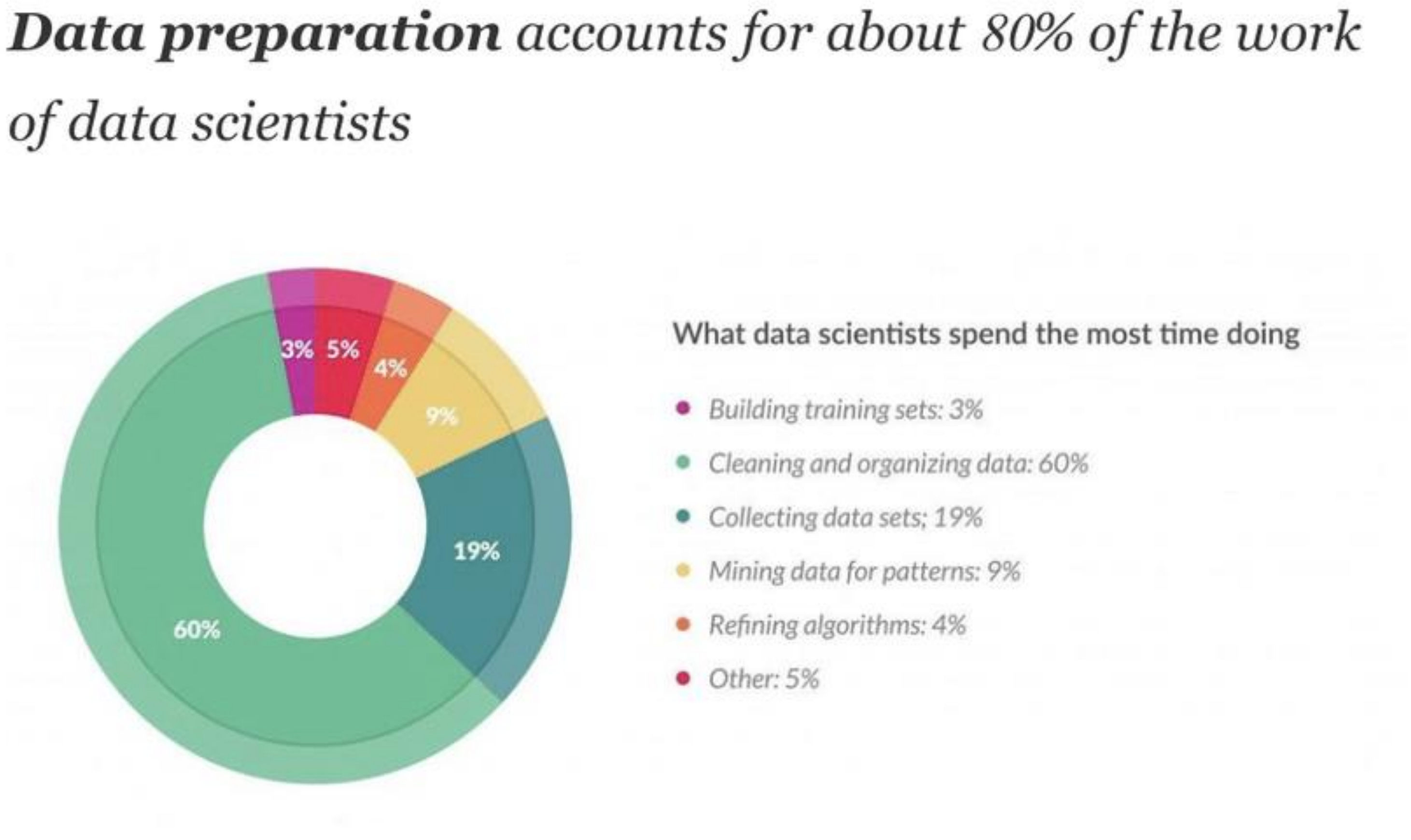
Week 6

Module 4 – Data Modeling

Overview

- Revisiting the data analysis pipeline
- Anatomy of a good figure
- Data distributions, linear distributions
- Fitting and interpreting linear models

Data wrangling can often be a large component of the total analysis



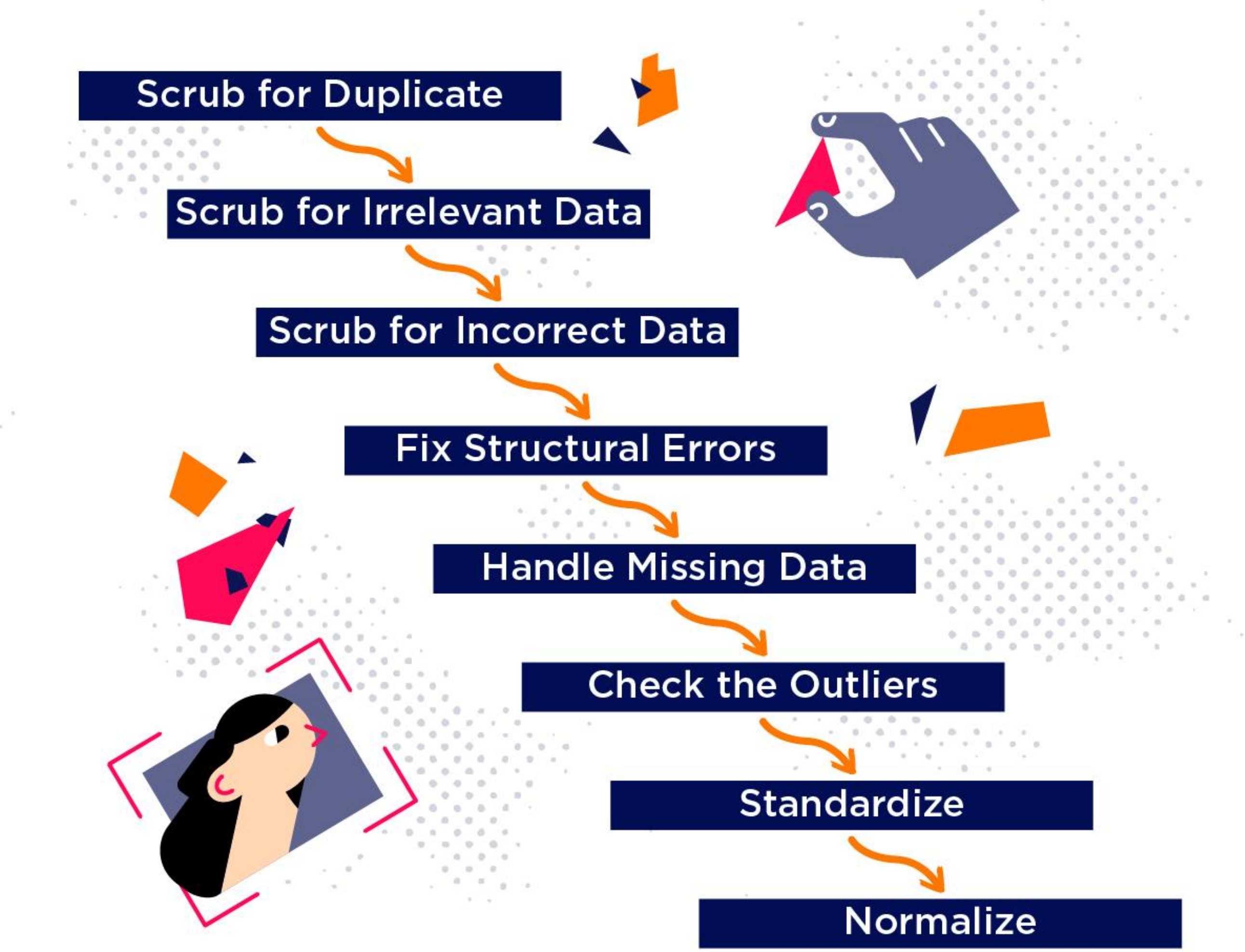
***Data preparation** accounts for about 80% of the work of data scientists*

What data scientists spend the most time doing

- ● *Building training sets: 3%*
- ● *Cleaning and organizing data: 60%*
- ● *Collecting data sets; 19%*
- ● *Mining data for patterns: 9%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

"Playing the whole game": A data collection and analysis exercise with Google Calendar

What to look for in your dataset



Scrub for Duplicate

Scrub for Irrelevant Data

Scrub for Incorrect Data

Fix Structural Errors

Handle Missing Data

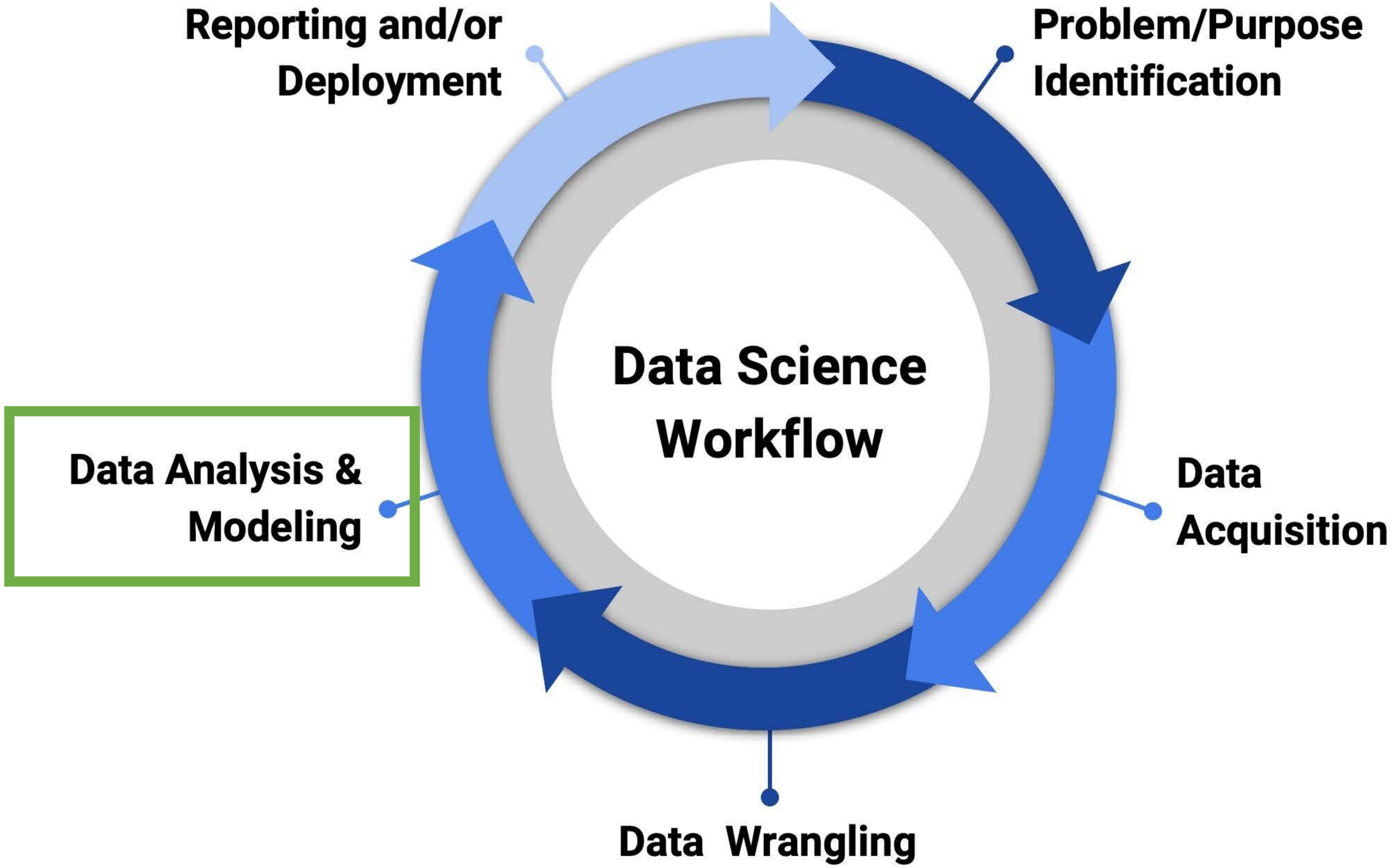
Check the Outliers

Standardize

Normalize

https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/

Next Step!


Reporting and/or Deployment
Problem/Purpose Identification
Data Science Workflow
Data Acquisition
Data Wrangling
Data Analysis & Modeling

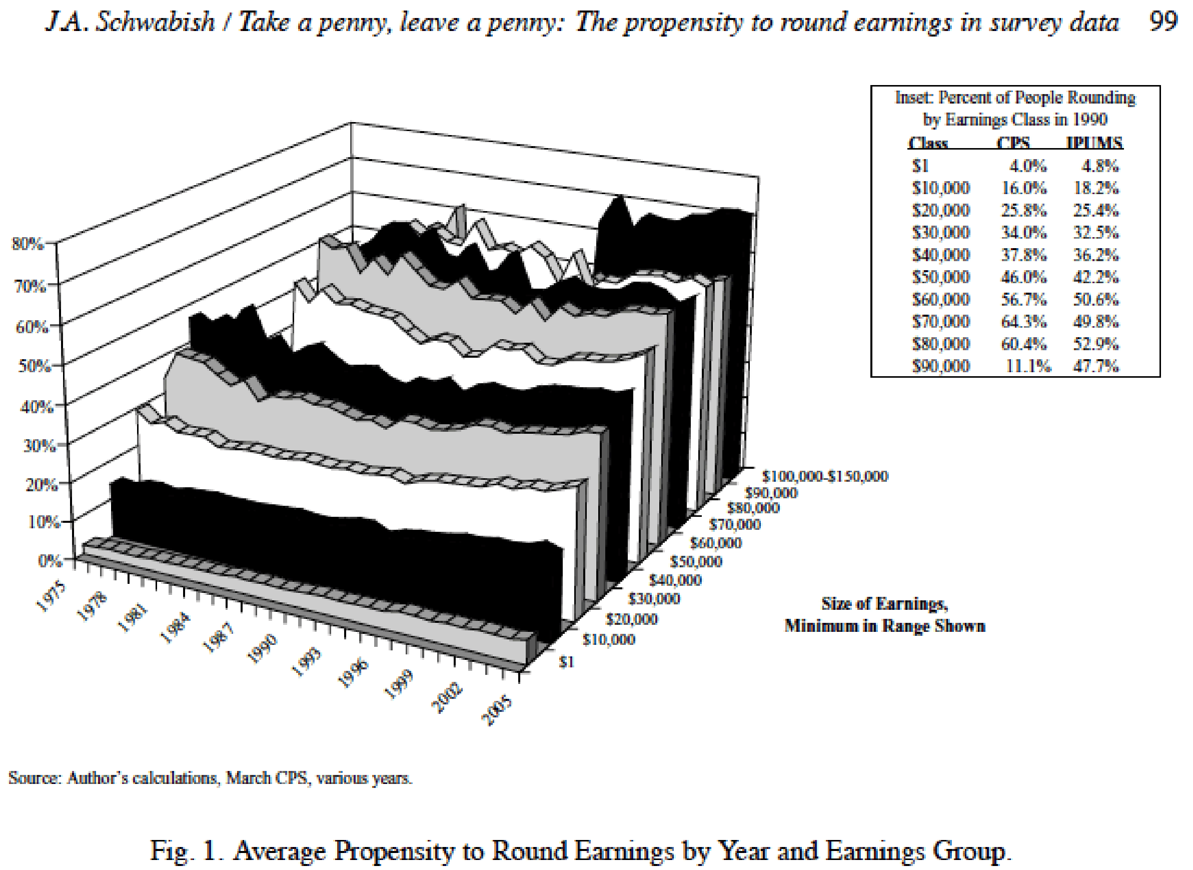
Always start with data wrangling

- Need to start with tidy and clean data before analyzing

- Take some time to understand the data before you can identify what needs to be corrected

"Garbage in, garbage out"
Your analysis is as good as your data.

What do we like or dislike about this plot?



J.A. Schwabish / Take a penny, leave a penny: The propensity to round earnings in survey data 99

Inset: Percent of People Rounding by Earnings Class in 1990

Class	CPS	IPUMS
$1	4.0%	4.8%
$10,000	16.0%	18.2%
$20,000	25.8%	25.4%
$30,000	34.0%	32.5%
$40,000	37.8%	36.2%
$50,000	46.0%	42.2%
$60,000	56.7%	50.6%
$70,000	64.3%	49.8%
$80,000	60.4%	52.9%
$90,000	11.1%	47.7%

Source: Author's calculations, March CPS, various years.

Fig. 1. Average Propensity to Round Earnings by Year and Earnings Group.

What do we like or dislike about this plot?

Like

- One nicely labeled axis
- There is a figure legend

Like to improve

- Raw data is not necessary in the main figure of a plot
- What do the colours mean?
- Label the other axis

What do we like or dislike about this plot

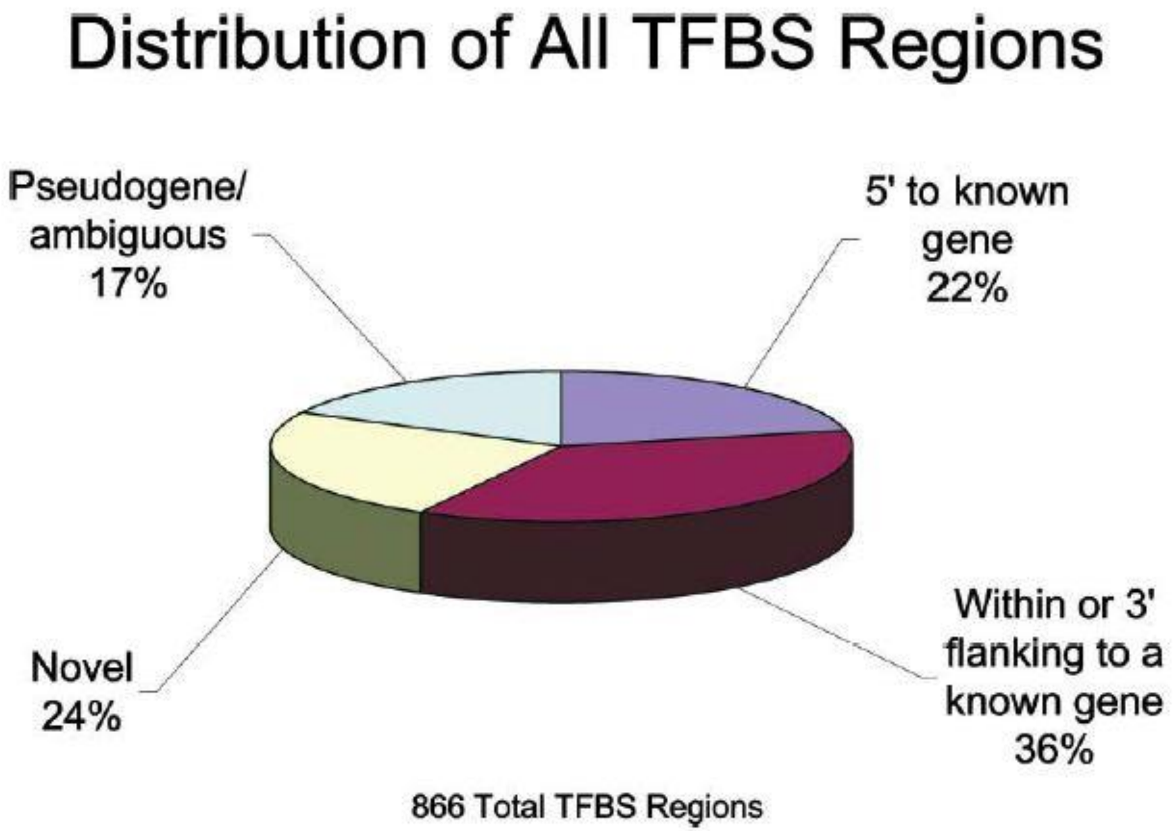


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

What do we like or dislike about this plot?

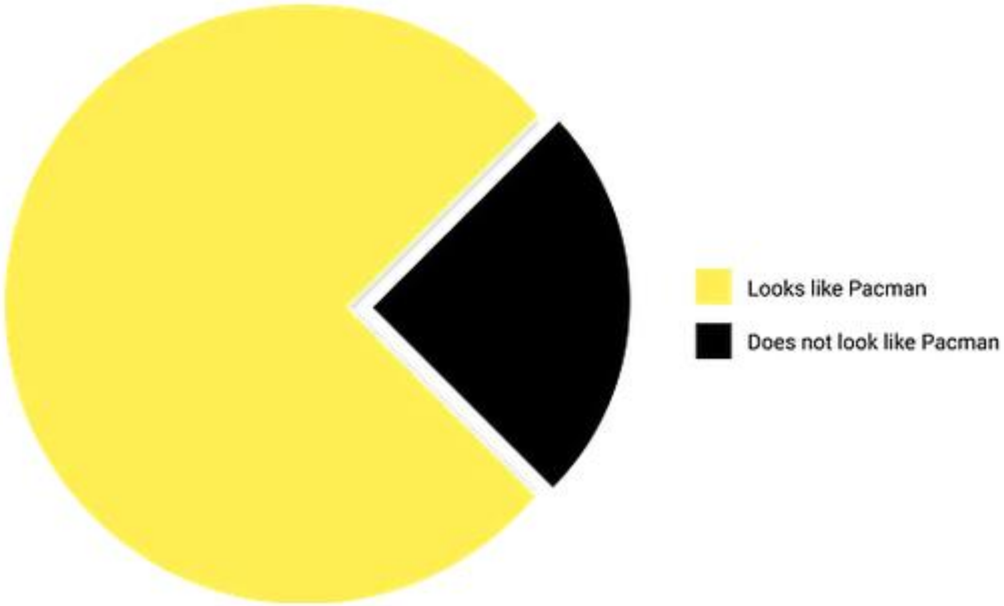
Like

- Slices are clearly labelled
- Explanation in the figure legend is helpful

Like to improve

- Pie charts make it difficult to compare the groups
  - 22% vs 24%
- 3D often adds nothing but confusion to interpreting a plot

One of the few good reasons to use a Pie Chart



Exploratory Data Analysis

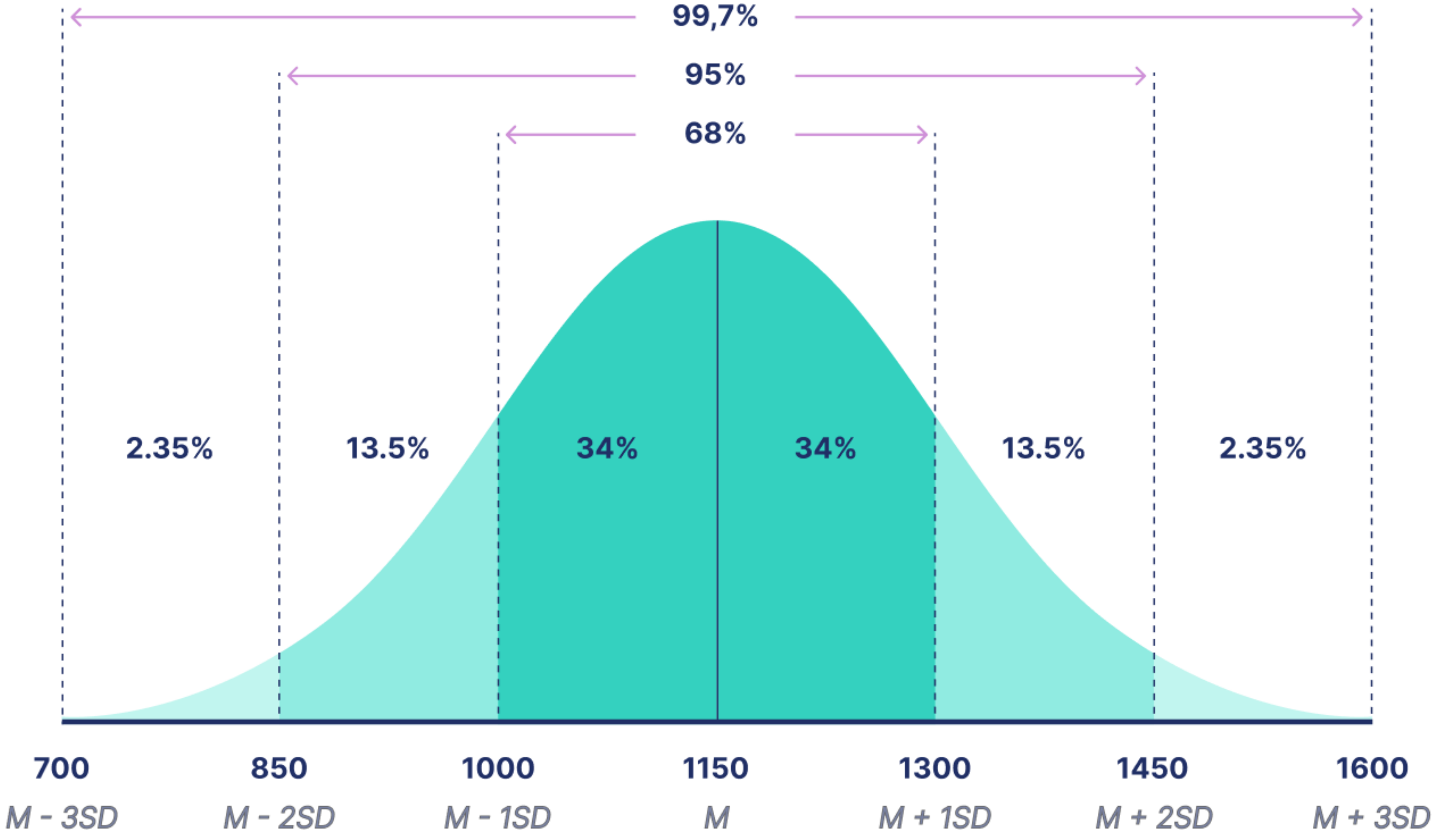
- Looking for trends or patterns in the data
  - Can be newly collected data or re-analyzing published/existing datasets

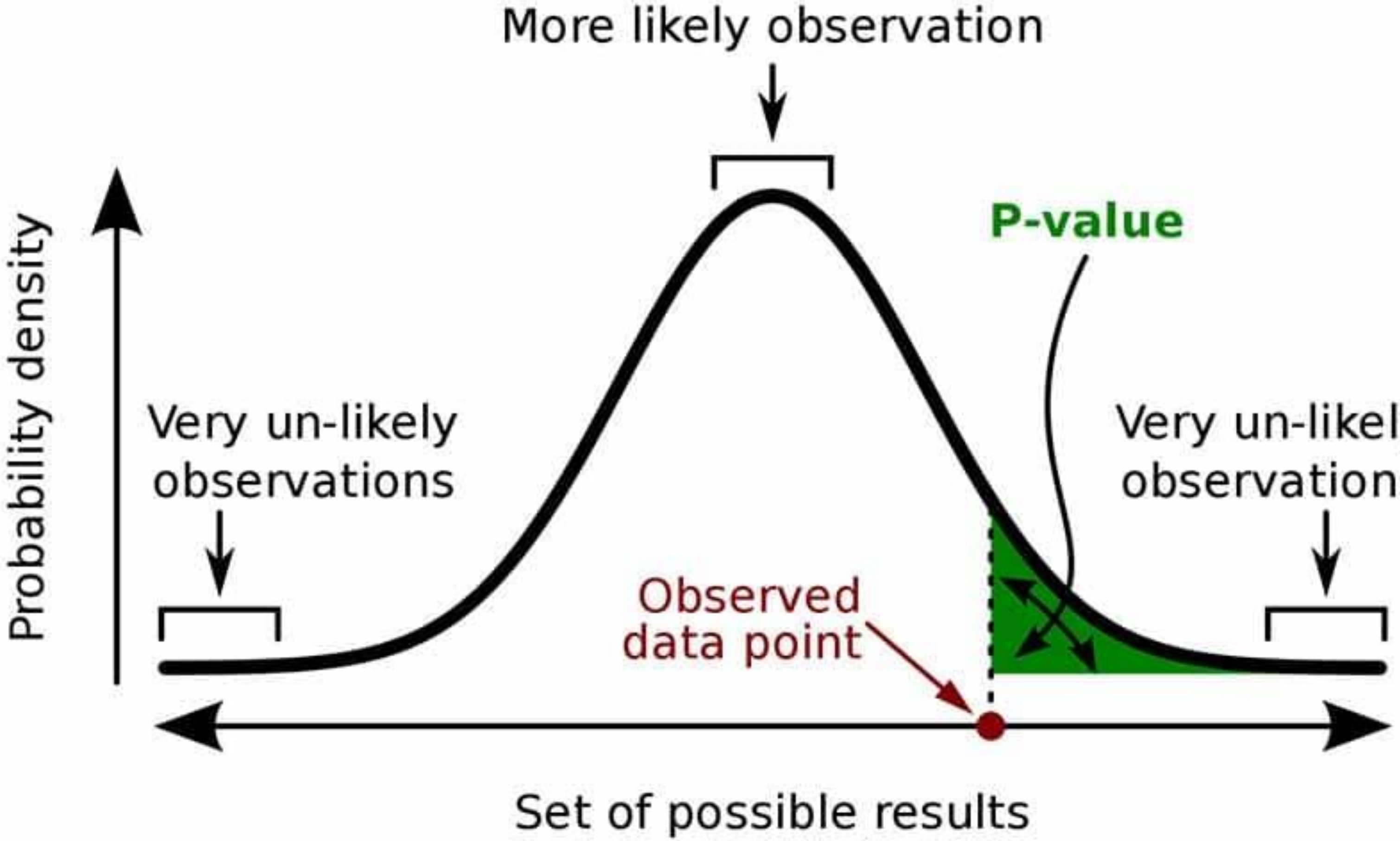
- Can be difficult like data wrangling
  - No workflow is universal to all datasets
  - Much of the troubleshooting work will not make it into the main figures of a publication
  - Lots of trial and error

Integrating all the tools

- Introduction to R
  - Data types, basic syntax
- Data wrangling
  - Principles of tidy data
  - Reshaping and indexing data with base R and tidyverse
- Visualization
  - Investigate patterns and relationships in numeric and categorical data

Normal (“Gaussian”) Distribution

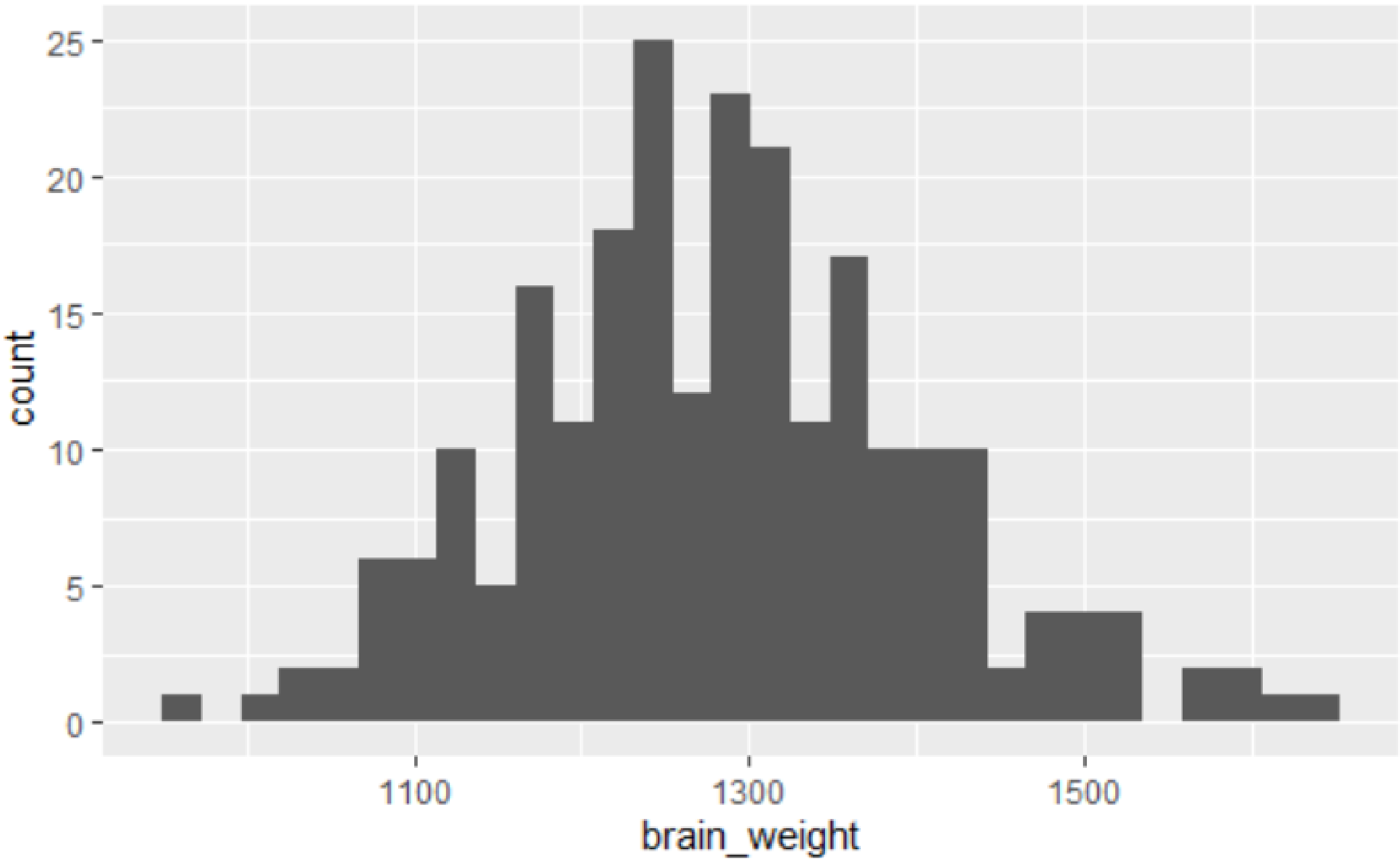




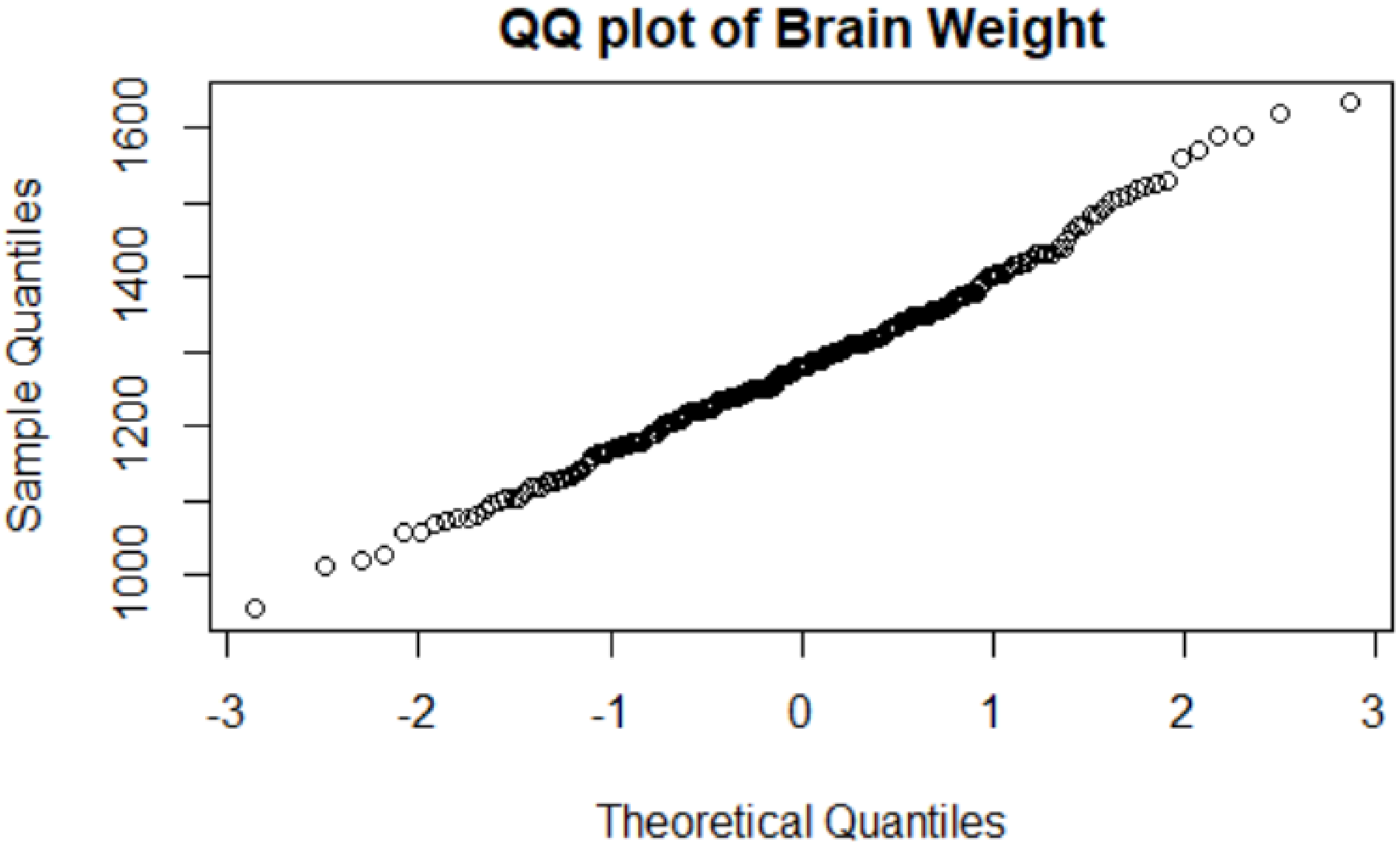
A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Continuous Variables

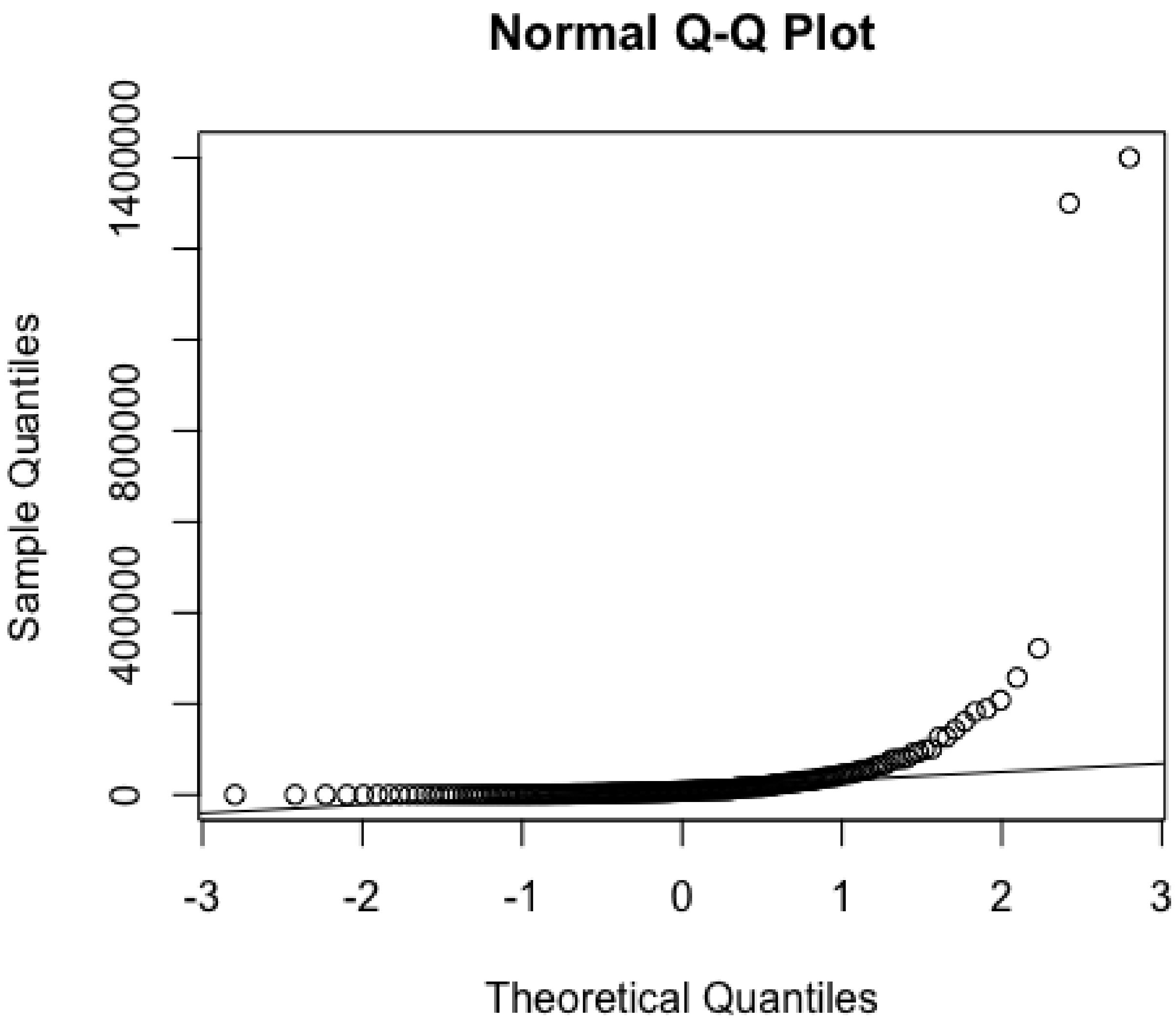
Does this variable follow a normal distribution?



Continuous Variables



Continuous Variables



Simple linear model

- Always between two continuous variables

Two purposes:

- Confirm a relationship between the variables
- Predict the value of the dependent variable using the independent variable

Simple linear model

$$x \rightarrow y$$

- Predictor
- Independent variable

- Output
- Dependent variable

$$y = mx + b$$

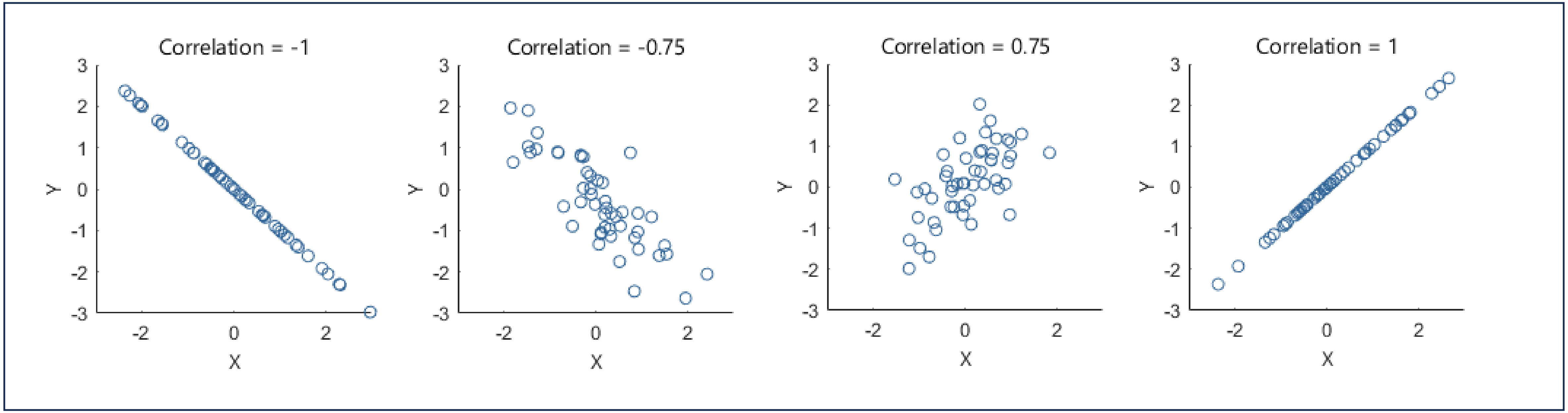
$$y = mx + b$$

Slope

- Modify the coefficient, one per coefficient

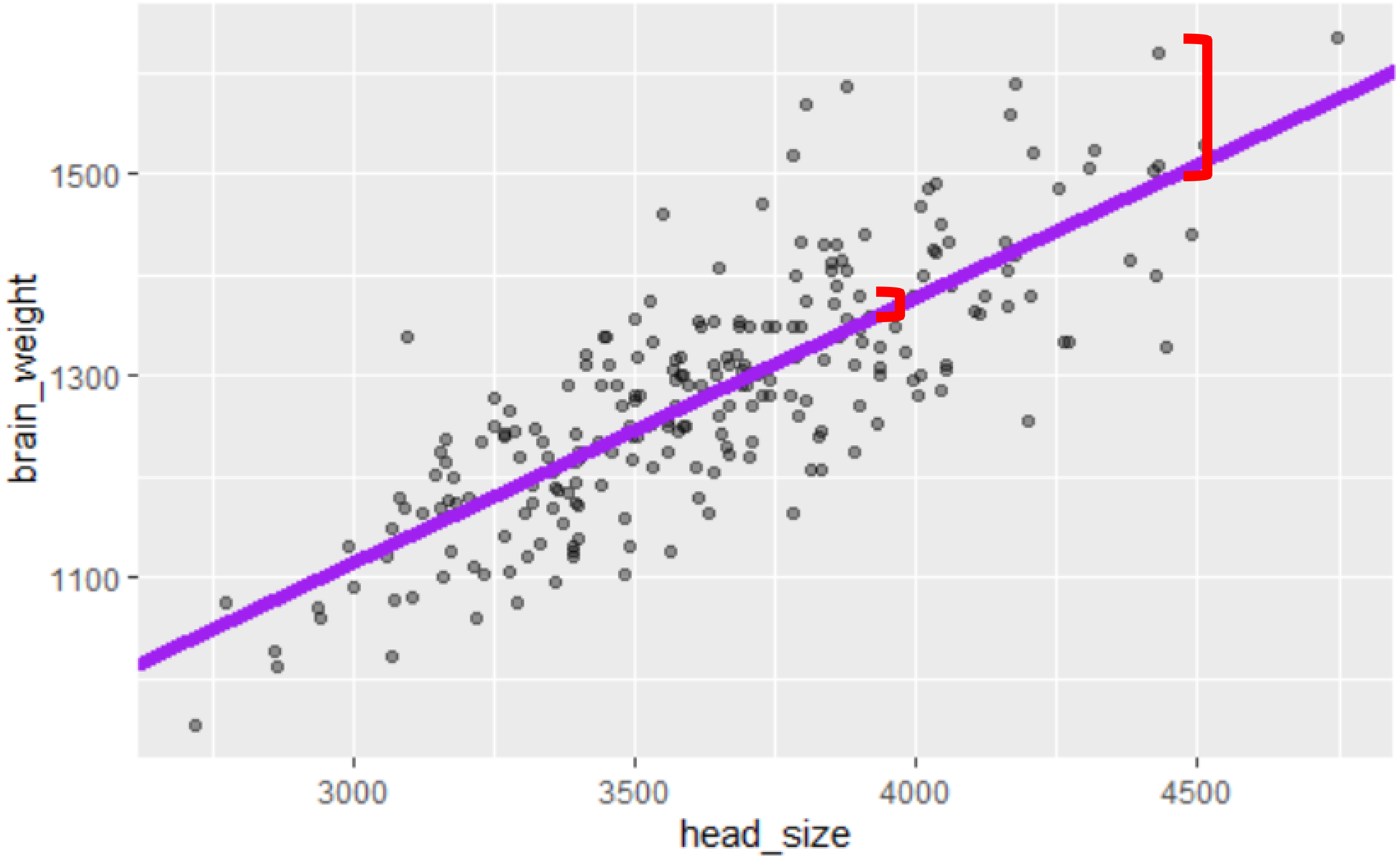
Intercept

- One per equation



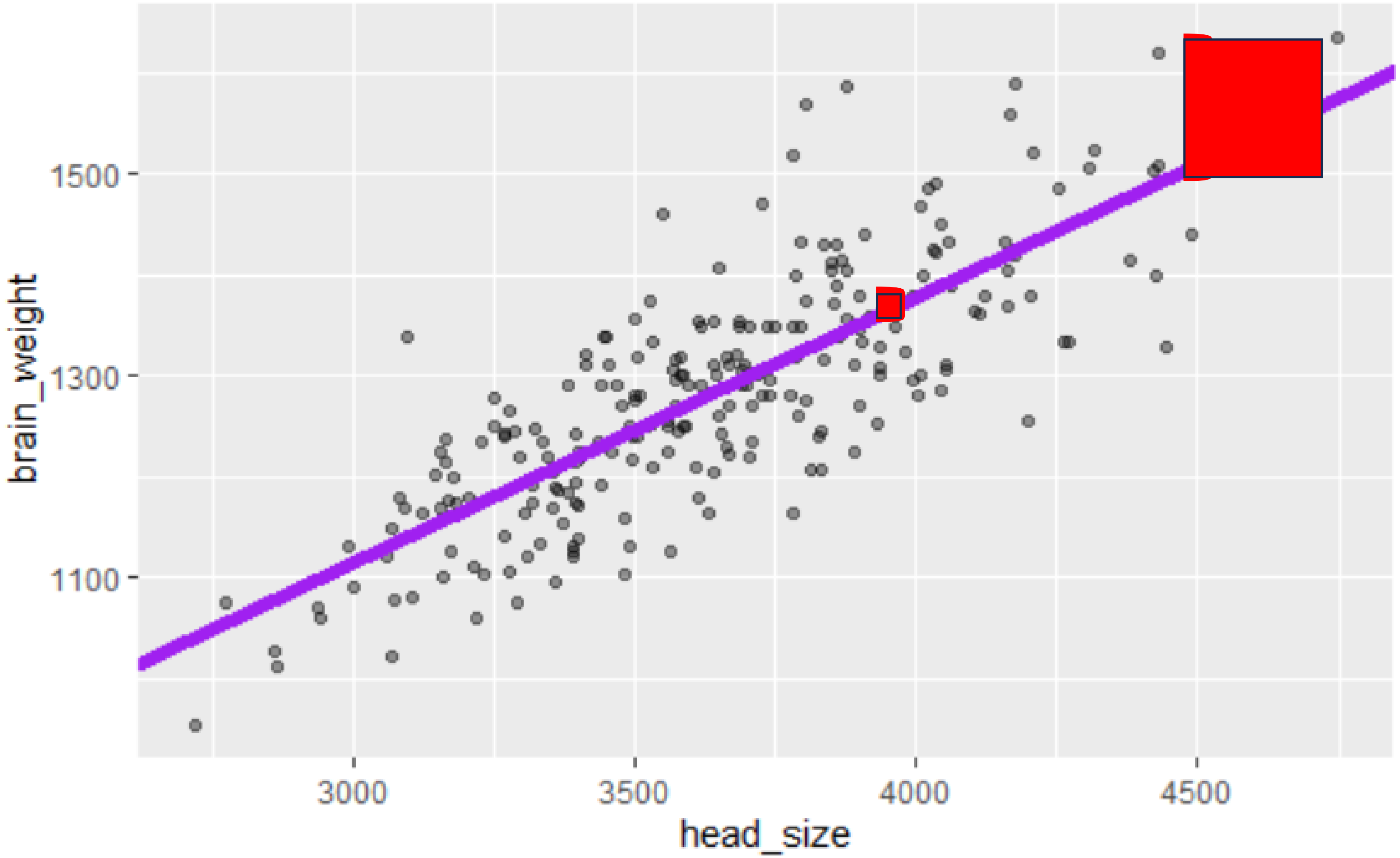
Fitting a linear model

Minimize residuals (difference between the observed and expected values)



Fitting a linear model

Minimize residuals (difference between the observed and expected values)



Assumptions of a linear regression

1. The X and Y variables are connected by a linear relationship

2. The residuals are independent (independence)

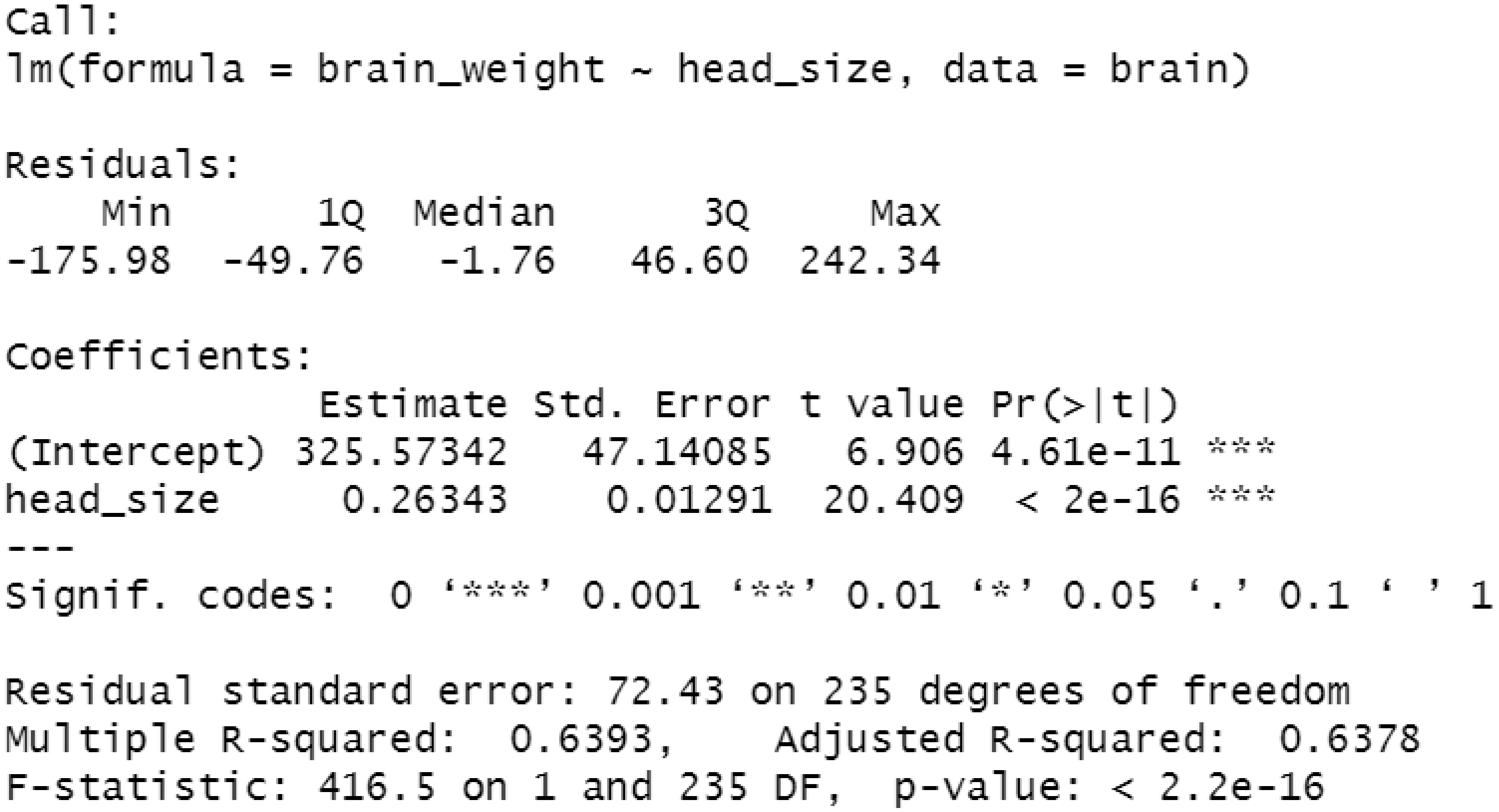
3. At every value of the independent variable, the residual is constant (homoscedasticity)

4. The residuals of the model are normally distributed (normality)

https://www.statology.org/linear-regression-assumptions/

Using R

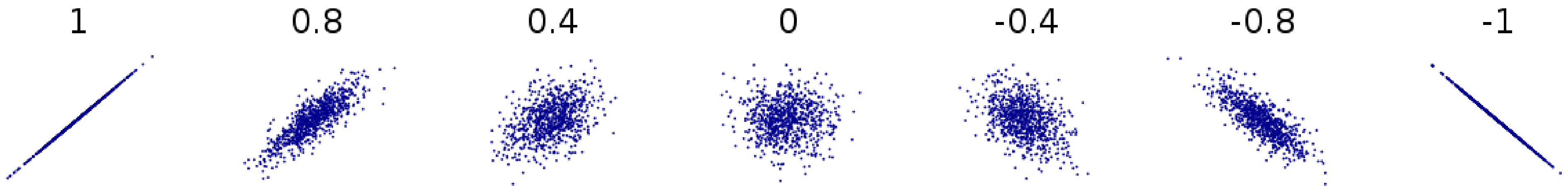
- Running the code is simple
- Interpreting the output is the challenge here!



```
Call:
lm(formula = brain_weight ~ head_size, data = brain)

Residuals:
    Min      1Q  Median      3Q     Max 
-175.98  -49.76   -1.76   46.60  242.34 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 325.57342   47.14085   6.906 4.61e-11 ***
head_size     0.26343    0.01291  20.409  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72.43 on 235 degrees of freedom
Multiple R-squared:  0.6393,	Adjusted R-squared:  0.6378 
F-statistic: 416.5 on 1 and 235 DF,  p-value: < 2.2e-16
```

$$y\ (brain\ weight) = (0.263)\big(x\ (head\ size)\big) + 325.573$$

Look at both the numbers and the visuals



Wrap Up

- Start with high quality data and remember to clean/tidy your data before starting the analysis

- Anatomy of a “good” figure
    - Choose the right format for your data, never use 3D plots

- Statistical tools rely on assumptions that must be met by the dataset in order for the results to be valid

- Simple linear models can only model data that follows normal distribution