Exploratory Data Analysis

UBC R Study Group

Aaron Quinton

Oct. 31, 2018

Acknowledgements/References

- Online Text *R for Data Science* by Garrett Grolemund & Hadley Wickham
  R for Data Science Ch. 7 - EDA
- The *Hass Avocado Board* and Justin Kiggins for compiling Avacado data
  Dataset csv download link on Kaggle
- Data Visualization published in *Nature Methods*
  Nature.com Blog
- DataCamp Course by Rick Scavetta: Data Visualization with ggplot

Intro to Exploratory Data Analysis (EDA)

EDA is a creative process with the purpose of **exploring** your data to generate quality questions. It is not quantitative but more qualitative in nature and is iterative as the exploration inspires questions which become more exploration and so on.

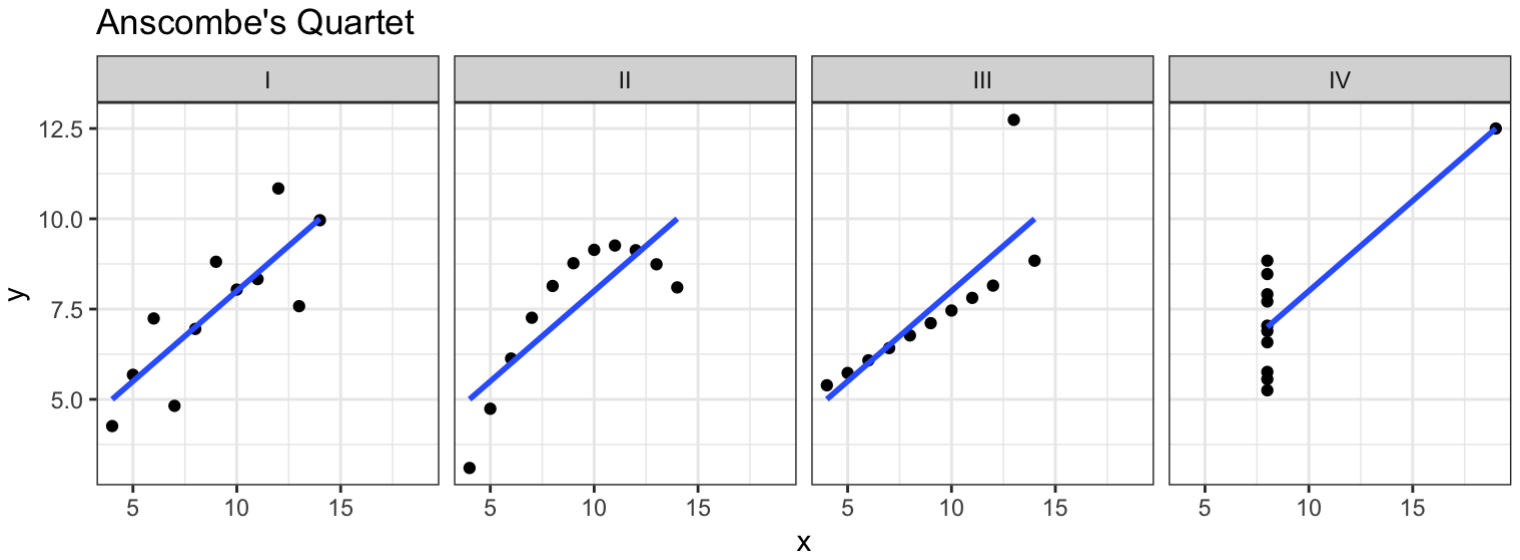
“There are no routine statistical questions, only questionable statistical routines.” — Sir David Cox

A good starting place for questions:

- What type of variation occurs within my variables?
- What type of covariation occurs between my variables?

To answer these question its crucial to know if our data is Categorical or Numerical.

The Motivation behind Visual EDA



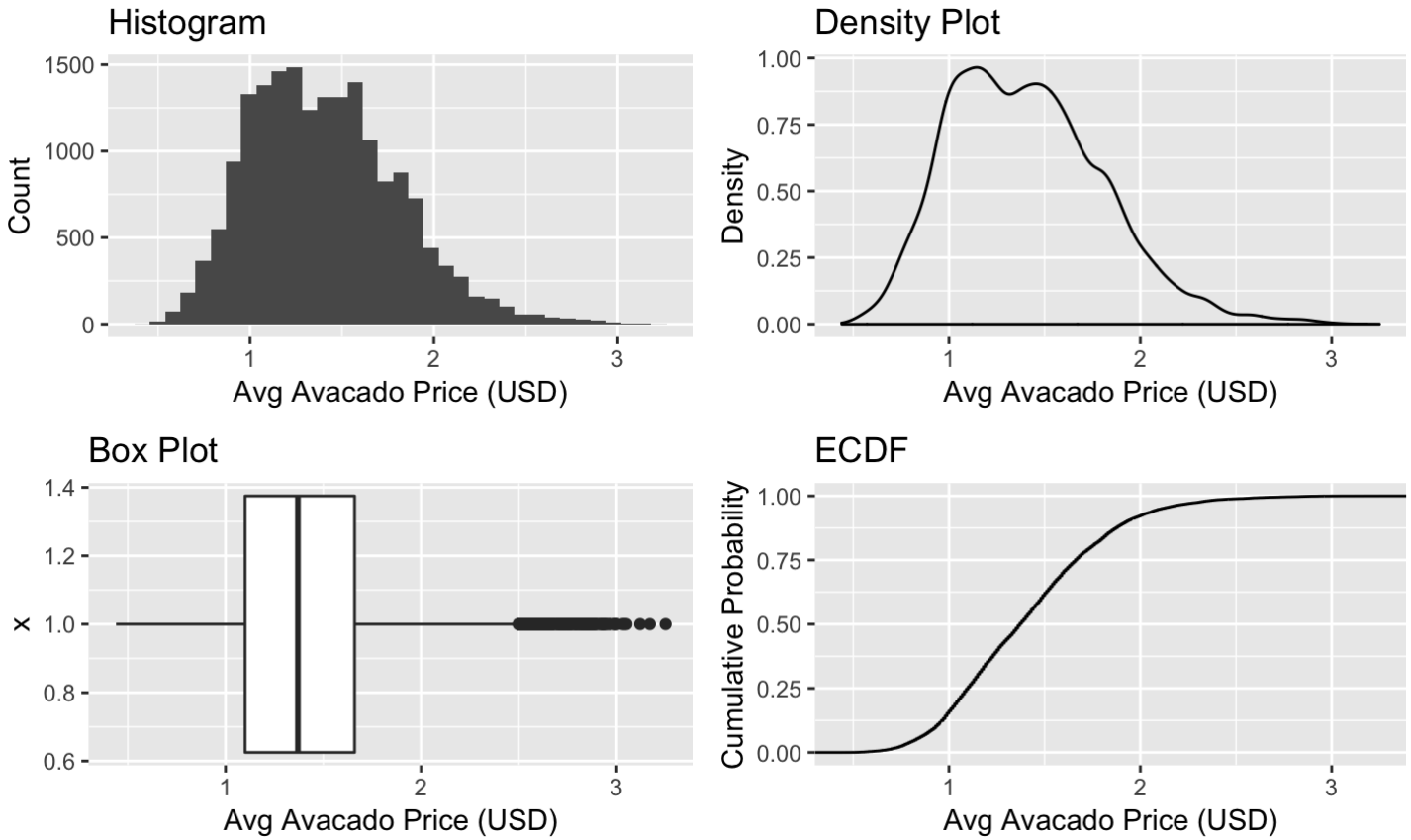

```
## # A tibble: 4 x 6
##   set   `mean(x)` `sd(x)` `mean(y)` `sd(y)` `cor(x, y)`
##   <fct>     <dbl>   <dbl>     <dbl>   <dbl>       <dbl>
## 1 I             9    3.32      7.50    2.03       0.816
## 2 II            9    3.32      7.50    2.03       0.816
## 3 III           9    3.32      7.5     2.03       0.816
## 4 IV            9    3.32      7.50    2.03       0.817
```

Before Getting Our Hands Dirty

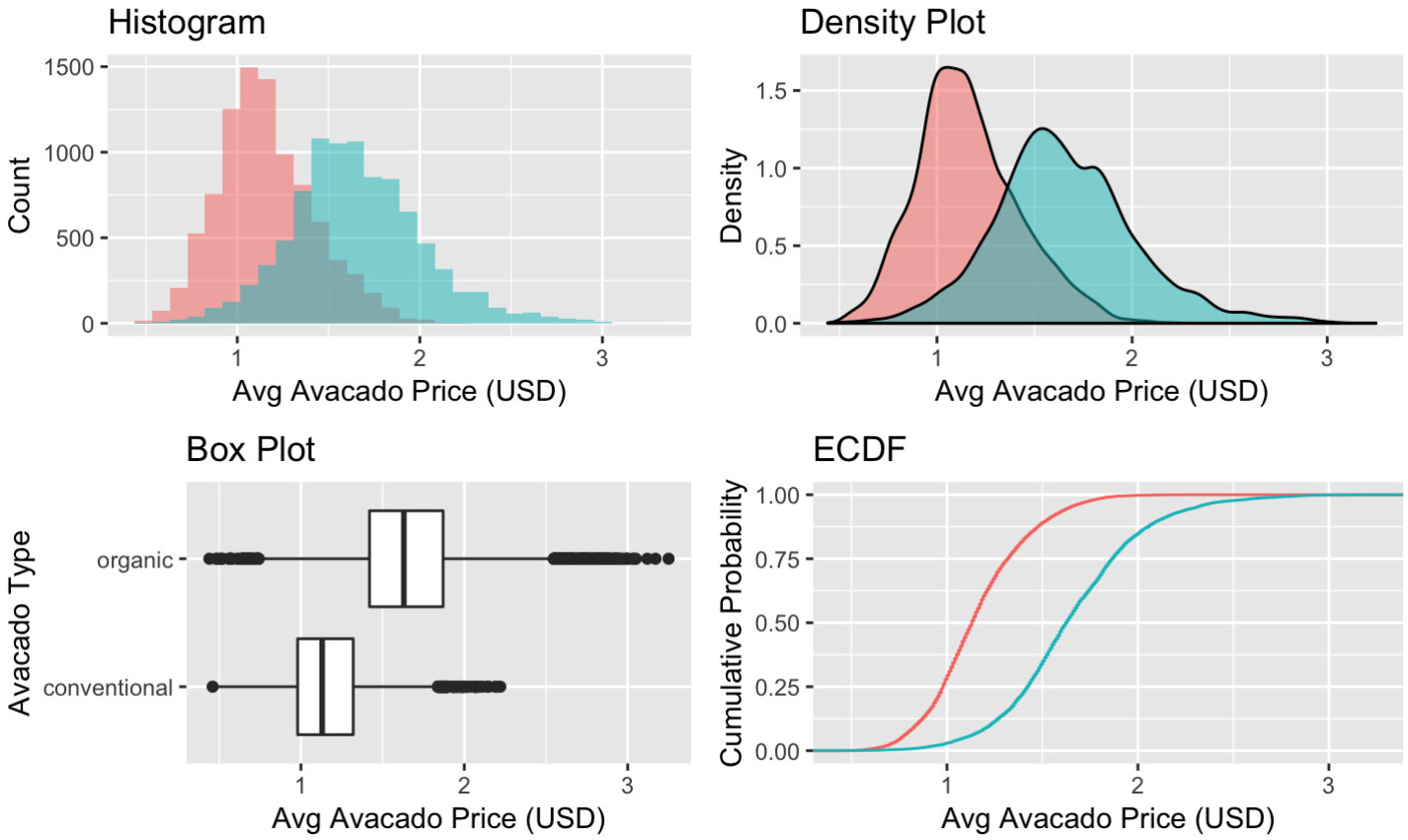
Finding Data

- Google Dataset Tool Search
- Kaggle Datasets
- Awesome-Public-Dataset Repo
- Built in R Datasets or Dataset packages on CRAN. To see the list type `data()` in the console.

Exploring One Numerical Variable

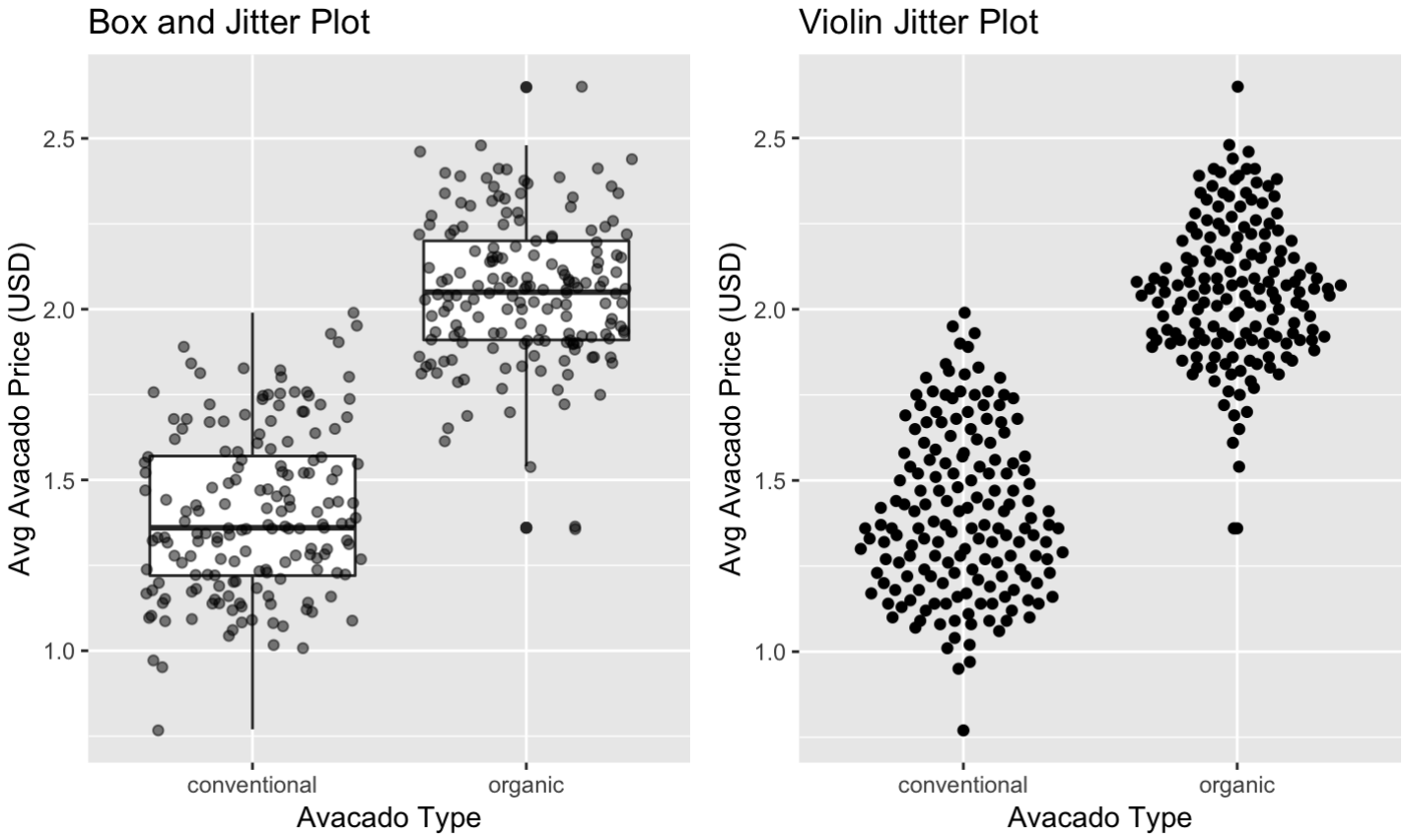


One Numerical One Categorical

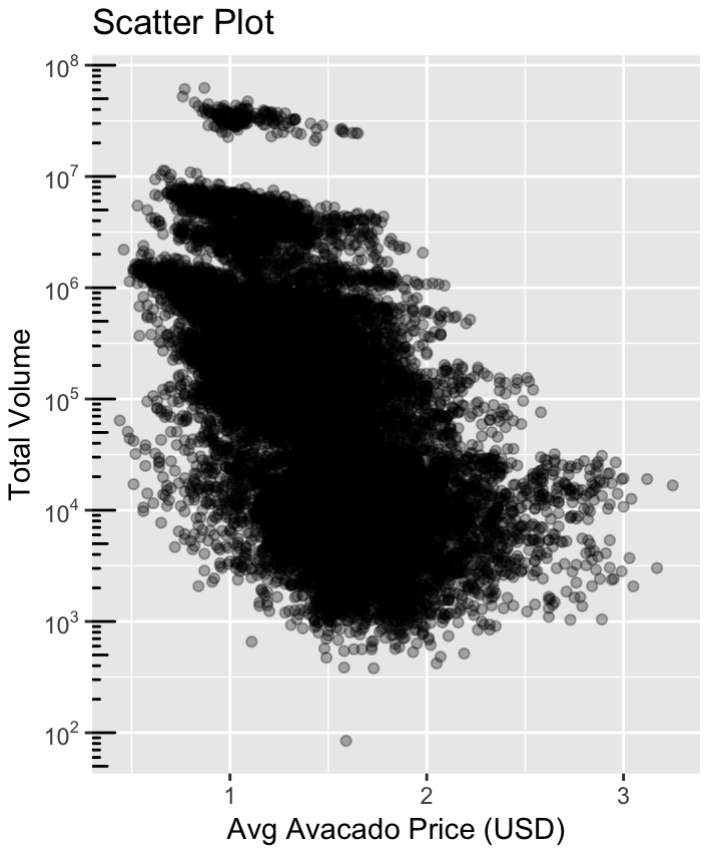


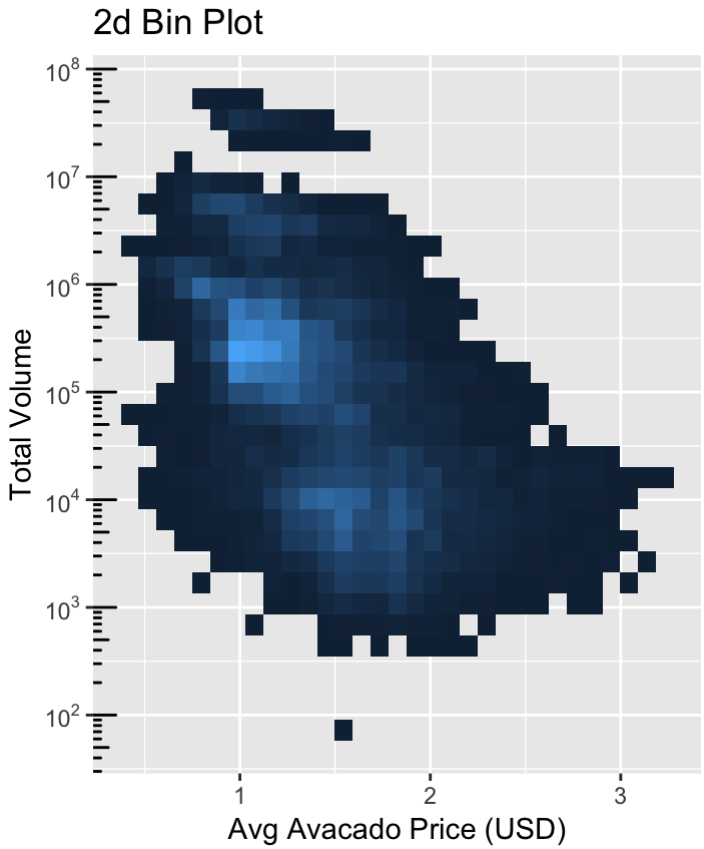
One Numerical One Categorical (Cont.)

Using `geom_jitter()` and `geom_quasirandom()` with the data set filtered to the region New York.



Two Numerical Variables





Exploring More than Two Variables

So far we have been primarily been using position mapping and/or color mapping. But we are able to map more variables to the following aesthetics:

Aesthetic	Description
x	X axis position
y	Y axis position
color	Color of dots, outlines of other shapes
fill	Fill color
size	Diameter of points, thickness of lines
alpha	Transparency
linetype	Line dash pattern
labels	Text on a plot or axes
shape	Shape

Categorical Visual Guide

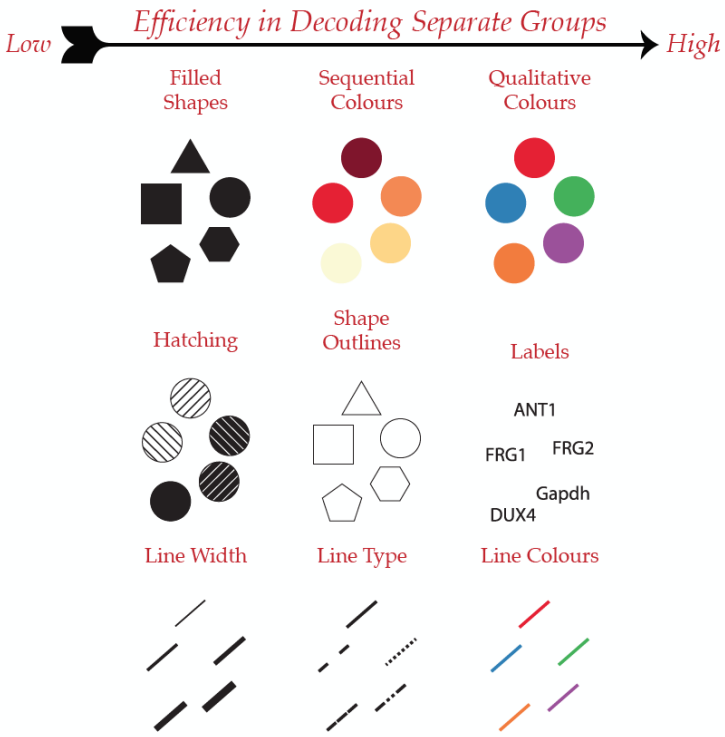


Image adapted by DataCamp from Wong,B, *Nat Met*, 7(9), 2010, p665

Continous Visual Guide

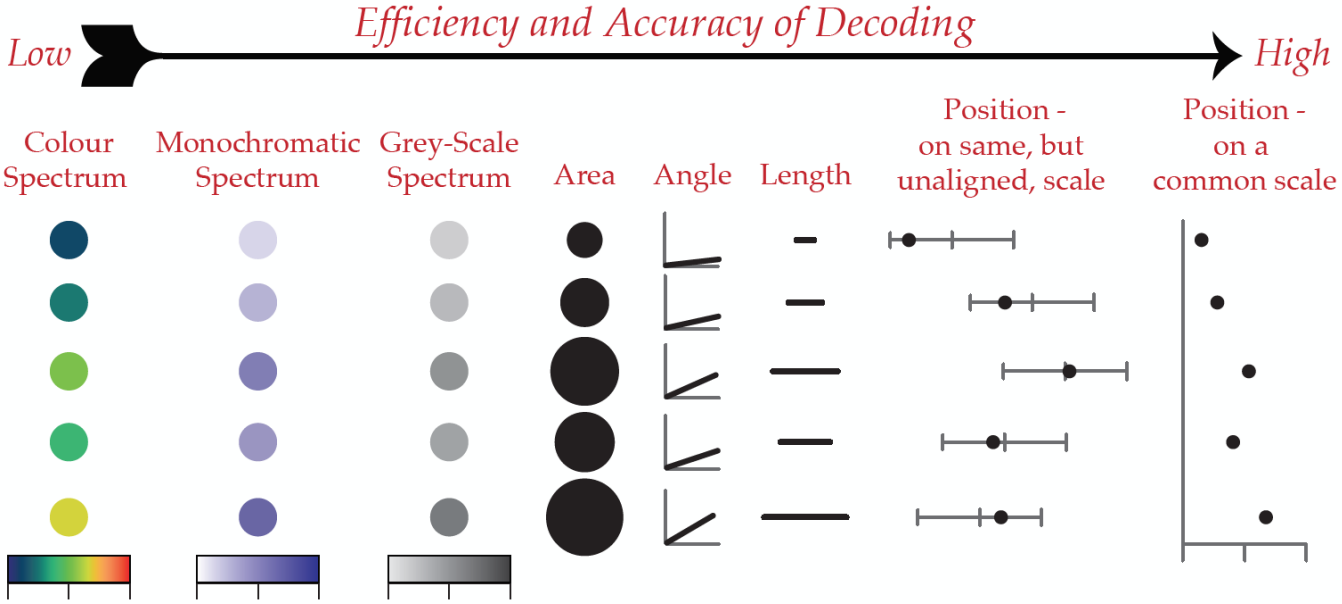


Image adapted by DataCamp from Wong,B, *Nat Met*, 7(9), 2010, p665