DATA ANALYSIS IN R - WHAT TO COVER?

MEASURES AND CALCULATION	VS	WHAT TO DO ABOUT IT?

MEASURES OF CENTRAL TENDENCY

CALCULATE THE MEAN VALUE.
CALCULATE THE MEDIAN.
CALCULATE THE MODE.

PRESENT ALL THREE VALUES TOGETHER.

MEASURES OF VARIABILITY AND DISPERSION

CALCULATE THE RANGE.
CALCULATE THE STANDARD DEVIATION (STDEV).
CALCULATE THE INTERQUARTILE RANGE (IQR).

PRESENT STDEV, USE IQR FOR OUTLIER DETECTION.

MEASURES OF SKEWNESS AND KURTOSIS

1. MEAN > MEDIAN
   I. NEGATIVE SKEWNESS
   II. EXPECT MORE VALUES ON HIGHER SIDE)
2. MEDIAN > MEAN
   A. POSITIVE SKEWNESS
   B. EXPECT MORE VALUES ON LOWER SIDE)

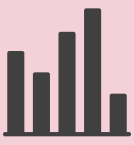
1. KURTOSIS = 0
   A. NORMAL DISTRIBUTION, MODERATE TAILS
2. KURTOSIS > 0
   A. HEAVIER TAILS
   B. SHARPER PEAK
   C. MORE OUTLIERS PRESENT
3. KURTOSIS <0
   A. LIGHTER TAILS
   B. FLATTER PEAK
   C. LESS OUTLIERS

VISUALIZATION IN R

USE:

- BOX PLOT
  - IDENTIFIES OUTLIERS
- SCATTER PLOT
  - SHOWS RELATIONSHIP BETWEEN TWO VARIABLES
- BAR PLOT
- HISTOGRAM
  - IDENTIFIES OUTLIERS
  - SHOWS THE DISTRIBUTION SHAPE

OUTLIERS? WHERE?!



VISUALIZE YOUR DATA. BOXPLOT AND HISTOGRAM CAN HELP TO SEE OUTLIERS.

ACKNOWLEDGE THEM, AND DON'T RUSH TO DELETE THEM, AS NEW OUTLIERS MIGHT ARISE BY DOING SO.

Cheese, Butter, and Data: A Complete R Data Analysis Guide using Dutch Dairy Production Data

This blog post is part of a "Statistics Essentials" series of stories about the basics of statistics, and its vocabulary. To read all the posts that are a part of it, click here. The list is freque...

Majanalytics / Sep 9

CHECK FOR FULL ARTICLE HERE!