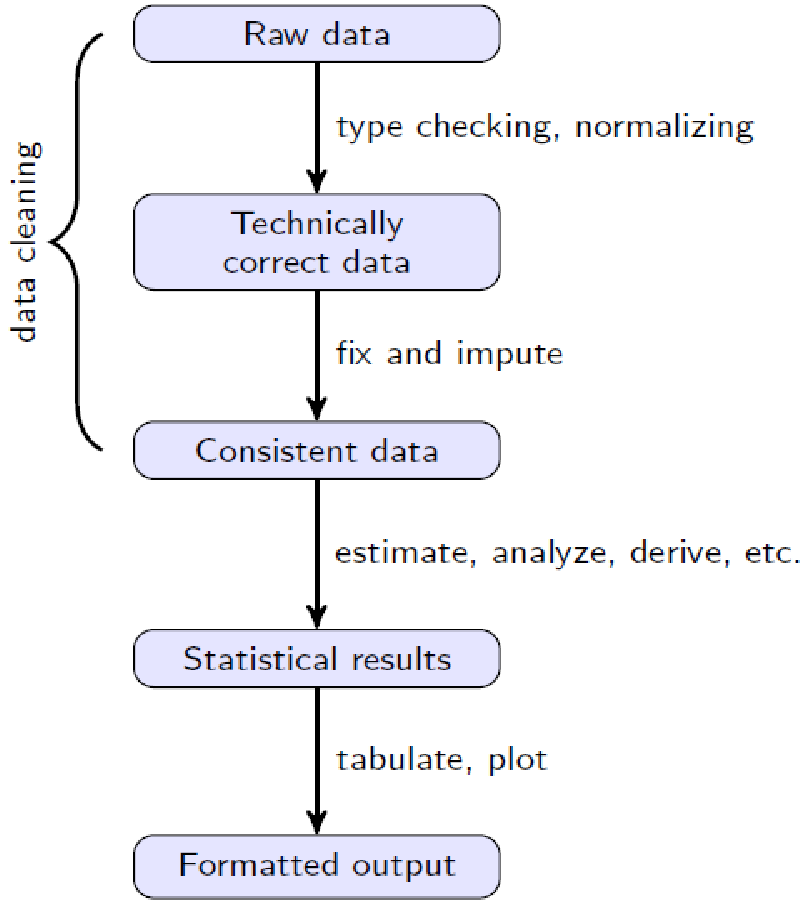
Descriptive Statistics: Applications

Pavan Kumar A
Senior Project Engineer
Big Data Analytics Team
CDAC-KP

Introduction To Data Cleaning

- Data Wrangling is the process of transforming raw data into consistent data that can be analyzed.
- Data cleaning is one of the primary pain points of data science.
- Data Scientists spend 80% of data analysis time in cleaning data.[1]



[1] http://www.crowdflower.com/blog/2014/01/data-cleaning-with-crowdflower-the-80-percent-solution-for-data-scientists

Source: https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

Raw Data

- Raw data can be hard to understand, even for those with advanced technical skills.
- In order to make this data easily understandable and user-friendly, it must be pre-processed and prepared for actual analysis.
- **Causes of Poor data quality**
  - Data entry errors
  - False values for variables
  - Heaping data
  - Application errors or Coding errors
  - Incomplete or outdated data
  - Differences in data representation among data sources
- **Problems associated with dirty data**
  - Invalid reports resulting in wrong interpretation

Steps: Data Cleaning

- Data cleaning is basically done in two steps DETECTION and CORRECTION.
- Some of them includes following
  - Missing data coded as "999”
  - The 'not applicable' or 'blank' coded as "0"
  - Reduplication
  - COLUMN SHIFT - data for one variable column was entered under the adjacent column
  - Logic checks
- Support of Domain expert is also needed for data cleaning.

Error Detection

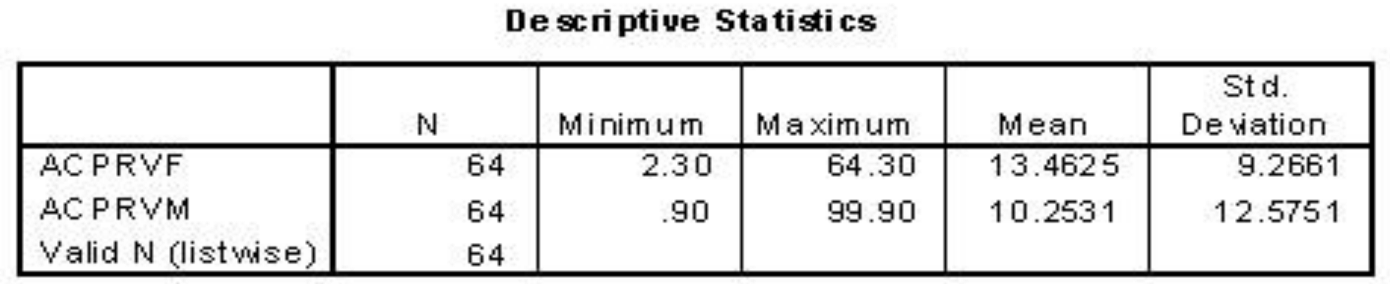
- Most of the errors will be detected using **Descriptive Statistics**
- **Descriptive Statistics are of three types**
  - Summary Statistics
  - Tabular Statistics
  - Graphical Statistics
- Summary Statistics
  - Min and Max
  - Mean
  - Median
  - Variance
  - SD (Standard Deviation)

Error Detection

Descriptive Statistics : Summary Analysis

- Look at minimum and maximum values (range) for descriptive statistics
- Look for Likeliness of the value in terms of range or z-score
- Look at Mean, Median and Standard Deviation

- Example 1:



Descriptive Statistics

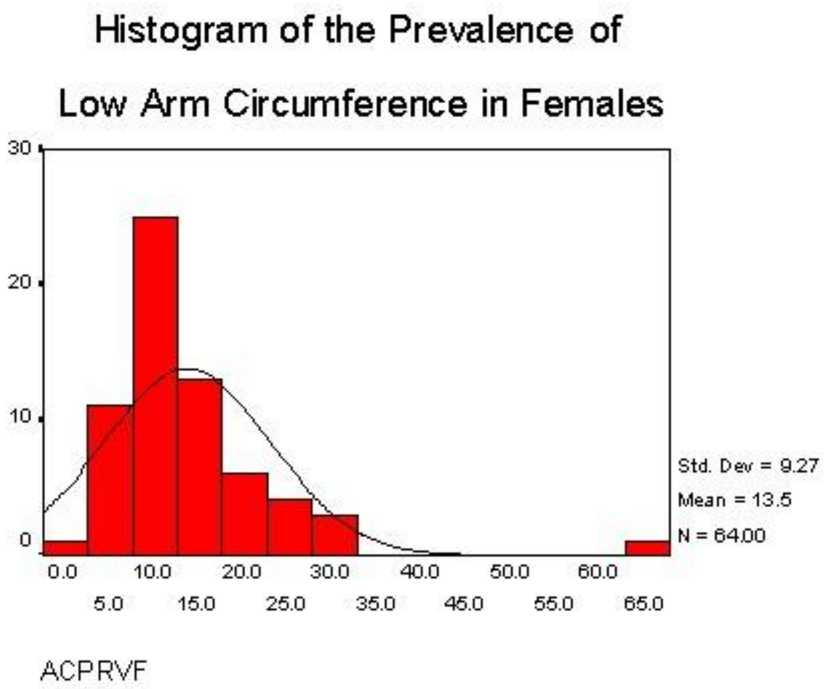
	N	Minimum	Maximum	Mean	Std. Deviation
ACPRVF	64	2.30	64.30	13.4625	9.2661
ACPRVM	64	.90	99.90	10.2531	12.5751
Valid N (listwise)	64				

Source: http://www.tulane.edu/~panda2/Analysis2/datclean/stats_with_errors.html

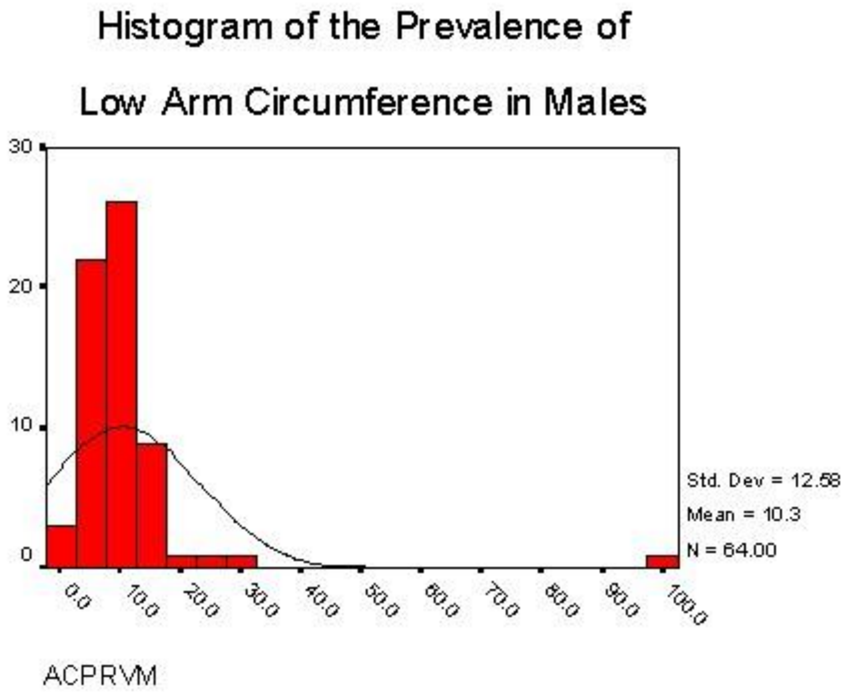
- **ACPRVF:** Females low arm circumference in cm’s (age<5 yrs)
- **ACPRVM:** Males low arm circumference in cm’s (age<5 yrs)

Error Detection

- **Descriptive Statistics : Graphical Analysis (Histogram)**



Histogram of the Prevalence of Low Arm Circumference in Females

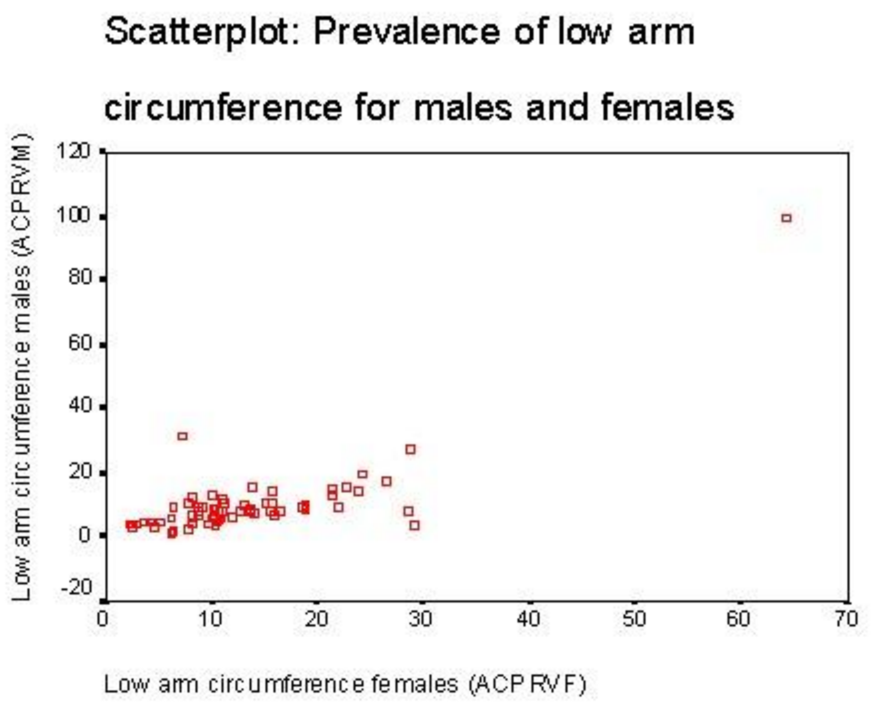


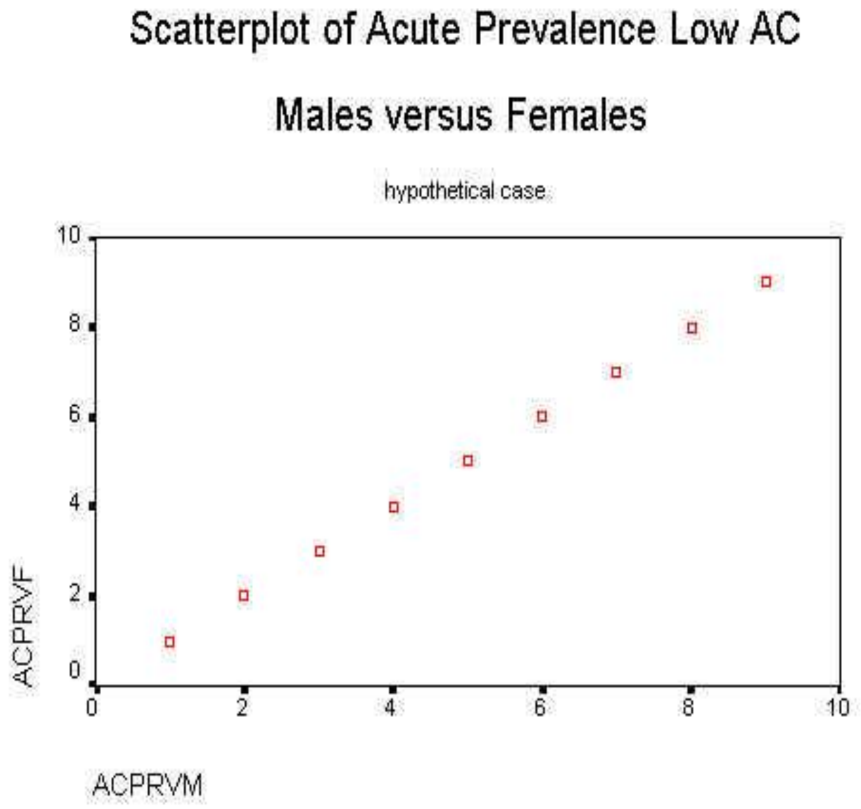
Histogram of the Prevalence of Low Arm Circumference in Males

Source: http://www.tulane.edu/~panda2/Analysis2/datclean/stats_with_errors.html

Error Detection

- **Descriptive Statistics : Graphical Analysis (Scatter Plot)**
- Some errors appears only when it is compared with two variables.
- Outliers are one of those to look at.

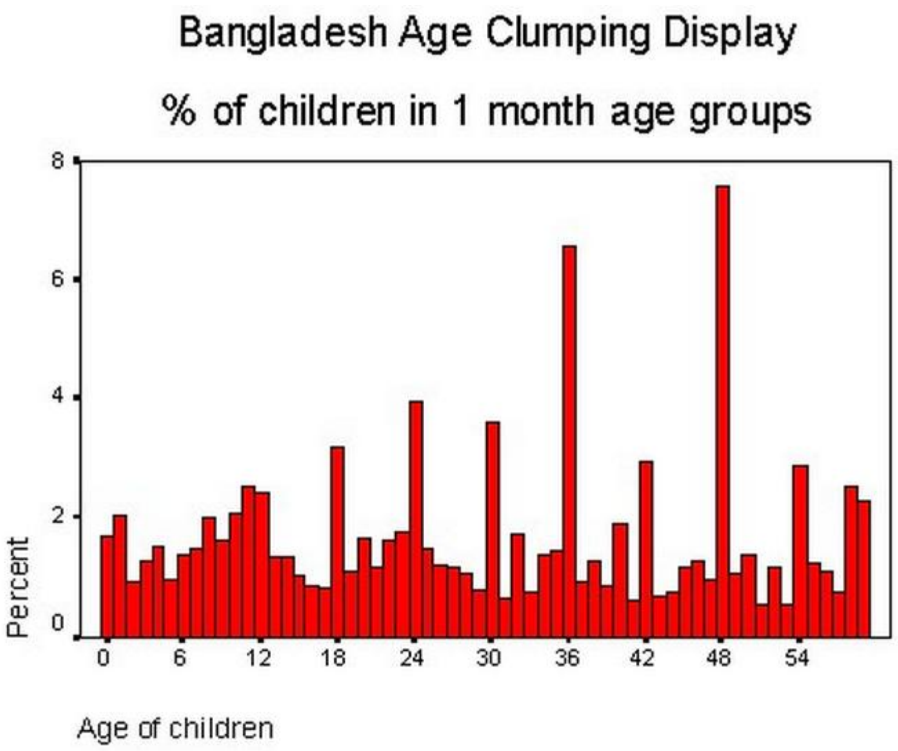




Source: http://www.tulane.edu/~panda2/Analysis2/datclean/stats_with_errors.html

Error Detection

- **Descriptive Statistics : Tabular Analysis (Frequency)**
  - Frequencies help to locate the 'dirty' data (Unequal distribution) among the entered variables.
  - Example 2: Baby ages

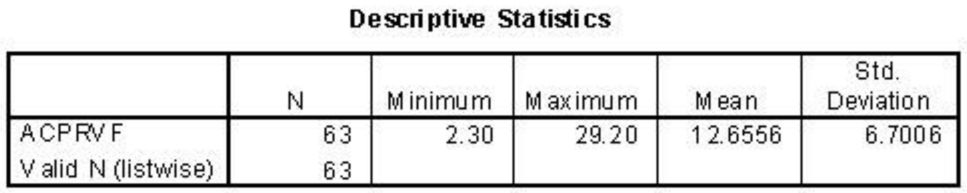


Error Detection

- **Logic Checks**
  - We can often detect errors in data simply by seeing if the responses are logical.
  - Example
    - We would expect to see 100% of responses, not 110%.
    - Issuing driving license for the age group <18

Error Correction

1. Categorize the values like <=60% and >=60%-100% and assign the values 0 and 1 respectively. (This eliminates the unexpected ranges)
2. Outliers set to “missing” if the errors are very less



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ACPRVF	63	2.30	29.20	12.6556	6.7006
Valid N (listwise)	63				

3. Best way: Outliers set to “MEAN” (for multiple variable analysis) for normal distribution of the data values.

THANK YOU !!!!