



## **Data Collection**

Data origin

Merging multiple sources

Question(s) of interest: can data support?

Domain knowledge



## Understanding the Data

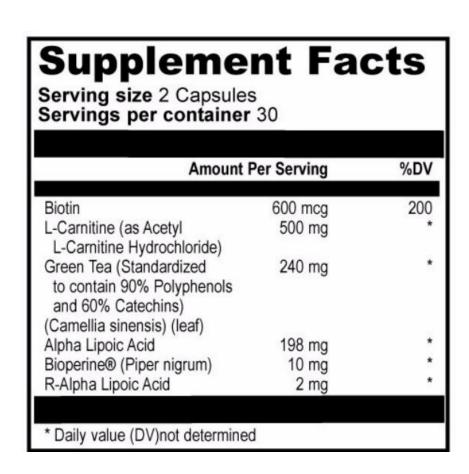
Meaning of each variable

Data formatting software reads correctly?

Ranges of variables

**Duplications** 

Outliers (errors?)



Use lots of graphics and summaries

## Data Preparation

Choice of variables

Choice of scales (continuous/categorical)

Binning and "unbinning"

## Missing values

Extent, type of missingness
Drop observations? Drop variables?
(replace with dummy?)
Impute (mean, regression, more advanced methods)
Explanatory vs. predictive

Creating derived variables

