Data Preparation

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?)

Supplement Facts

Serving size 2 Capsules
Servings per container 30

	Amount Per Serving	%DV
Biotin	600 mcg	200
L-Carnitine (as Acetyl L-Carnitine Hydrochloride)	500 mg	*
Green Tea (Standardized to contain 90% Polyphenols and 60% Catechins) (Camellia sinensis) (leaf)	240 mg	*
Alpha Lipoic Acid	198 mg	*
Bioperine® (Piper nigrum)	10 mg	*
R-Alpha Lipoic Acid	2 mg	*

* Daily value (DV)not determined

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