SESSION 6

DATA WRANGLING

Making your data ready for analysis…

HERTIE'S DATA SOCIETY

AJE JOSHUA A.

28.10.2019

TABLE OF CONTENTS

01

WHAT IS DATA WRANGLING?

02

FIVE STEPS & WHY?

What is Data Wrangling?

"…cleaning and organizing data into the desired format for future analysis by you or by others."

What do you when presented with messy data?

a) Refuse it
b) Accept and discard it
c) Get your coffee and start tidying it
d) Accept you are not cut out for data science

…but how do I tidy my data?

Not just how; also why?

INSPECT

Understand what is in your data in order to know how you want to analyze it

STRUCTURE

Organize the data e.g. turn a single column into several rows, for easier analysis

CLEAN

Correct wrongly imputed data, adjust skewed data, to improve data quality

ENRICH

Strategize about how your data might be augmented by additional data to enrich it

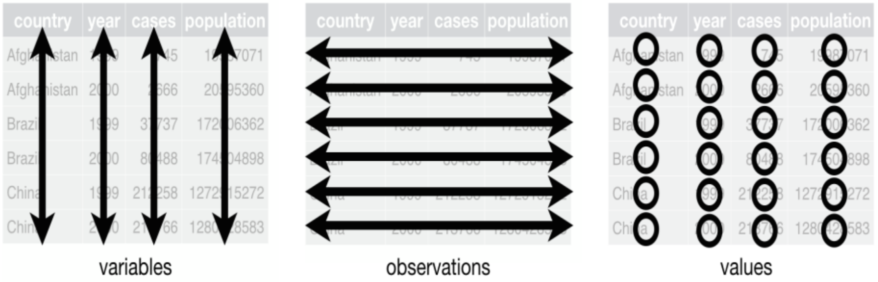
VALIDATE

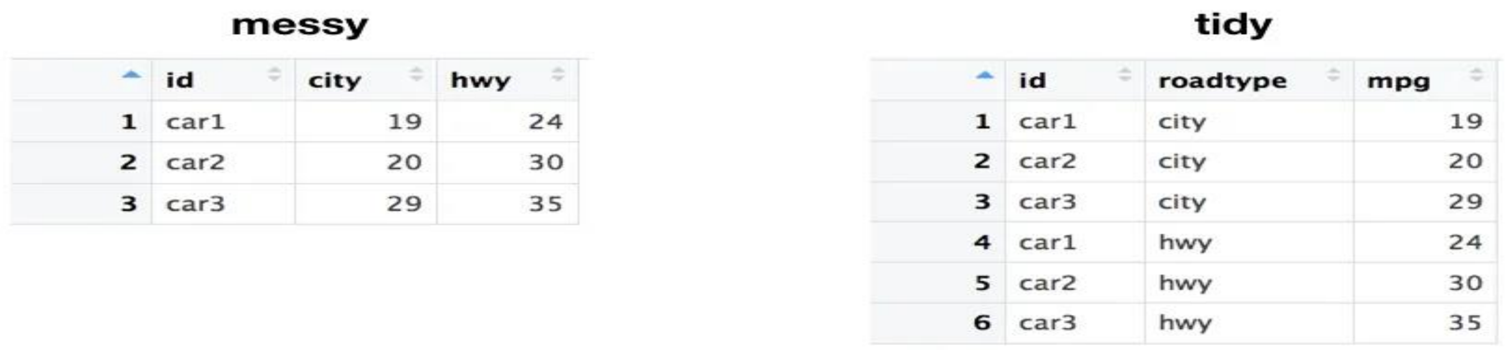
Check for accuracy of data fields to ensure data quality and consistency

Further insights: Data structuring

Three Golden Features of a tidy dataset:

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell





messy

	id	city	hwy
1	car1	19	24
2	car2	20	30
3	car3	29	35

tidy

	id	roadtype	mpg
1	car1	city	19
2	car2	city	20
3	car3	city	29
4	car1	hwy	24
5	car2	hwy	30
6	car3	hwy	35

Further insights: Data cleaning (missing values)

…values that should have been recorded but were not

Dealing with missing values:

1. **Deletion**
   - Suitable if amount of missing data is very small relatively to the size of the dataset
   - Simple method, but loss of information; could potentially wipe out all observations
2. **Mean/Median/Mode Imputation**
   - Generalized imputation, e.g. replace missing values with average of all non-missing values
   - Similar case imputation, e.g. calculate the average by gender and replace missing values based on gender
3. **Prediction Model**
   - Create a predictive model to estimate values that will substitute the missing values of a variable
   - If the variable with missing value has no relationship with other variables in the dataset, then the model will not be precise for estimating missing values
4. **Other means :**
   - kNN imputation
   - R Packages that deal with missing data: MICE, Amelia, Hmisc, missForest

Let’s practice!