COMP 333 — Week 7 Missing Values

Missing Values

The article on data cleaning has good coverage of missing values:

The Ultimate Guide to Data Cleaning: When the data is spewing garbage, by Omar Elgabry.

`https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4`

The ways to handle missing values are

- **drop** the observation with a missing value
- **drop** the feature/column with too many missing values
- **flag** the missing value with a new "value"
  such as pandas NaN (not-a-number),
  "Missing" for categorical variables,
  zero (0) for numerical values
  (which can be ambiguous as zero may be a legitimate value)
- **impute** (infer) a value for the missing value
  by simple or advanced inference techniques
    - **zero** for numeric variables
    - the **median**, **mean**, or **mode**, of the column values
    - a **random** value from the column values
      hot-deck imputation selects a random value
      from similar (nearby) column values
      after clustering observations using other column values
    - **interpolation** for numeric variables from similar (nearby) column values
      after clustering observations using other column values
      often using **linear regression**
    - **imputation**, inference, or prediction from similar observations
      using machine learning

Chapter 5 of the pandas book has a section on *Handling Missing Values*

Missing Values

Definition *Missing values* occur when no data value is stored for the variable in an observation.

Missing values are a common occurrence.

Missing values can have a significant effect on the analysis results from the data.

There are many causes of missing values:

- nonresponse in polls or surveys
- attrition of participants in longitudinal studies, or health studies
- errors in manual data entry
- inconsistency in data variables collected in different studies that are merged
- faulty instruments or sensors

Check the distribution of missing values:

Are they **random**, or not?

As this may help you find the cause behind the missing data.

Missing Value Imputation

Definition *Imputation* simply means replacing the missing values with an estimate, then analyzing the full data set as if the imputed values were actual observed values.

Advice for Handling Missing Values

For categorical values, you should **flag** them
typically with a value which as *"Missing"*

For continuous values of variable V, you should **flag** them

not with a value such as zero (0). which is ambiguous,

but by engineering a new feature/column *Missing_V*
that records T/F for whether the value in column V is missing or not.

Then you can select a method to impute the missing value in column V.

If a column does have too many missing values
then **drop** the column
or find the cause of the missing data and fix it.

If an observation has many missing values
then **drop** the observation
or find the cause of the missing data and fix it.

You should always try to find the cause behind messy data!

It is a good idea to **check the impact** of the missing values
and your approach to handling the missing values

by comparing the results of your analysis

using the treated missing values
versus removing all the missing values.