Data Cleaning Steps:

Cleaning is iterative, repetitive, and cyclic.

NOTE: This will *always* depend on the data and the format. **We will assume we are cleaning and preparing record** (rows and columns). The process would be very different for other data formats, such as transaction data, networks, image data, sequential, etc.

Cleaning Record Data:

1) First, consider each variable independently.
    a. What is its type – numerical, factor, integer, etc
        i. Why does this matter for analysis and for cleaning?
            1. For all supervised learning, the training label MUST be type factor.
        ii. Mathy analysis requires numeric data.
        iii. Be careful with integers – use numeric instead unless you AND UNDERSTAND why you need it to be int.
        iv. Gender as 0 and 1 should not be numeric – this is a factor.
    b. Update and correct all variable data types.
2) As you check each variable and adjust the TYPE, observe (visually) that variable, its name, and some of the data.
3) **<u>Missing Values:</u>**
    a. Blank or NA or NaN.
        i. In R, you can read in data to make sure that blank is converted to NA.
        ii. Try this on a small dataset to see how it works.
        iii. ALWAYS TRY EVERYTHING ON SMALL TOY DATA THAT YOU MAKE
            1. Try to count the NAs using R code.
            2. Is the count correct?

        STEPS:

        1) Discover
        2) Count
        3) Determine action – this will diff for diff variables

    ACTIONS for Missing values

    1) If 1% of less of your data size (number of rows) contains missing values, you may choose to simply remove those rows. This is a rare case.
    2) Sometimes a column has tons of missing values AND you do not really need it. In this case – you may choose to remove the column.
    3) In some cases, you will need to determine what to do with each variable.
        a. You will also need to look at the rows (which are the observations).

    OPTIONS:

    1) Removal – noted above.
    2) Replacement
        a. Using a measure – such as column mean, median, mode.

i. HUGE issues…
    1. Change the information in the data
    2. Change the variation
    3. It will smooth the data
    4. DO NOT replace with values like 0 or -1
    5. If your data is numeric – you may choose to replace with median BUT what if it is char/non-numeric
        a. You can use the mode – but this may not be OK

In some cases, you data will be labeled

1) Never replace (guess) at a label. Why??

Summary –

Missing data is a huge challenge. You do not want to lose data, but you must not alter the information contained in the data. Replacing or removing must be done with care AND you must keep note of the number of rows or columns removed, the percentage, and how it affect the mean, median, and variance if numeric, etc….TRANSPARANCY IN CLEANING

Incorrect Values – largely the same is cleaning missing values BUT they are harder to find.

1) For numeric – use and test ranges. Example: grades: check that all values are >=0 and <=100
2) Clustering
3) Visually – histograms, boxplots (outliers)
4) Observation – see if you can learn

Incorrect Format – but not incorrect data. Here you can correct it because it is not wrong.

Example: Fla rather than FL or Male rather then M or Virginia, VIRGINIA, Vigina

Challenge: finding these!

Outlier Detection and Correction

1) Visually: boxplots, hist
2) BE CAREFUL! Not all odd values are outliers.
3) NASA missed a planet or moon ☺
4) The hole in the ozone was missed for quite some time.

Visualize Every Variable:

1) Look at your data
2) Learn what information is in your data
3) Bars, pies, boxplots, violin plots, histograms, etc.
4) Look at the correlation between all pairs of variables – why does this matter?

Get Measures

1) Calculate the measures of center and variation for all numeric data.
2) Use **tables** to look at data

Reformat Data

1) Discretization (binning or categorizing)
2) Feature generation
3) Normalization