DATA 311 – Data Cleaning Thought Exercises

Names: _____ _____ _____

What would you do in each of the following situations?

LCD Weather data:

1. Hourly precipitation amounts contain "T" for "trace"; your goal is to compute total liquid precipitation

2. Hourly precipitation amounts contain "T" for "trace"; your goal is count total days it rained

3. All the Daily columns have NaNs when looking at rows that correspond to hourly observations

4. Suppose an observation for temperature simply wasn't taken at a small handful of hourly observations. Your goal is to plot temperature change over time.

NHANES (body measurements, etc.) or similar survey:

1. <2 year olds are missing standing height; your goal is to calculate the average height of the entire population, including babies

2. Left-handed vs right-handedness column is missing for .1% of entries in the dataset

Miscellaneous:

1. In a dataset of actors, there is no Year of Death for living people, but you wish to compute the average number films an actor plays in over their career.

2. In an assignment survey for for the first assignment of the quarter in one of my classes, one entry's Hours Spent column says 90.

3. In the Avengers dataset, a handful of Avengers are listed as having joined in 1900.