Lecture
slides https://docs.google.com/presentation/d/106TF1VtIrS3Ze09_9kgPZY14vPrc9wlMNg9E5XWVDo4/edit?usp=sharing

In almost every data science project, there is a large effort in organizing an analytic data set. This often requires data munging, web scraping, pulling data from a larger more complex dataset, merging datasets and formatting changes. In the ideal, this process goes very smoothly and the analytic dataset is a clean representation of the desired process. In real life, this process is fraught with errors.

As a manager, you likely won't be performing this effort, but rather managing it through reports during the process. How do you keep on top of data quality without being in the trenches? In this lecture, we give three strategies.

1. **The construction of summary tables**. The summary tables are very useful for catching data errors. Especially useful is keeping track of units and recording several summaries (means, medians, maxima, minima and other quantiles and standard deviations). By contrasting reports over time, you can check to see if things are changing in the processing that shouldn't be.
2. **Regression diagnostics.** Regression is a universal first step in analyzing data.Regression diagnostics are useful for catching data quality errors that manifest themselves in your analysis. Some useful regression diagnostics are:
3. **residuals** - the difference between the response and the fitted value, **hat diagonals, these** consider how variable a data row is among the space of predictors, **DF fits, DF betas, Cook's distance** these consider how much do fitted values and coefficients change when a point is not included in the fit? PRESS residuals, leave one out residuals - how much do predictions change when a point is left out of an analysis