Interview Questions: Data Analysis

Signal Data Science

- What is R^2? What are some other metrics that could be better than R^2?
  - R^2 represents the percentage of variation in the target variable explained by variation in the RMSE is a better measure of prediction accuracy than the square of the correlation.
- What is the curse of dimensionality?
  - Algorithms break down in high dimensions. The standard Euclidean distance metric doesn't work as well; distances between pairs of data points all become very similar.
  - See The Challenges of Clusternig High Dimensional Data by Steinbach, Ertoz, and Kumar.
- Is more data always better? gm
  - Yes from the perspective of pure prediction, but you may not always want to work with all the data you have (at least immediately); it's expensive to store a large amount of data, training models takes longer, the dataset may not fit in memory, etc.
- What are advantages of plotting your data before performing analysis?
  - See Anscombe's quartet – summary statistics don't tell all.
- How can you make sure that you don't analyze something that ends up meaningless?
  - Answer on Quora
  - Proper exploratory data analysis – graphing, looking at summary statistics, doing sanity checks on a lot of different hypotheses.
- How can you deal with missing values in your data?
  - Replace with mean/median/mode or use linear regression for multiple imputation. (Linearly regress each variable against the others).