KNN from *Doing Data Science* by O’Neil and Schutt plus notes from M. E. Waggoner

Process of K-Nearest Neighbors

1. Decide on your *similarity*  or *distance* metric
   1. *Distance* is not always calculated with the familiar Euclidean distance formula   
        .   
      For instance, that distance would not make sense if you were walking in New York City, because you can’t walk through walls. In that situation, the taxicab (or Manhattan) distance formula would make more sense:   
        .  
      There are many different ways to calculate distance, and the formula chosen will depend on the situation.
   2. *Possible distance or similarity measures*: Cosine similarity, Jaccard distance, Hamming distance, Manhattan distance, Euclidean distance, others.
   3. *Scaling* might have to be done to make the distance meaningful. There are long discussions on data analytics websites about when to scale and when not to. Generally scaling is done by normalizing the data so that each predictor has the same standard deviation, that is, by equalizing the spread of each variable.
2. Split the original labeled dataset into training and test data.
   1. You will split the data into training and test data sets so that you can test the process against known results.
   2. Cross-validation techniques discussed later in the book will help us create test and training sets.
3. Pick an evaluation metric.
   1. Misclassification rate is a good one. If you know the classification of the items in your test data, you can find the number and percent of those that are misclassified. This is the misclassification rate.
4. Run KNN a few times, changing *k* and checking the evaluation measure.
   1. There are ways to choose *k* to be odd, even, or other based on the characteristics of the data set.
5. Optimize *k* by picking the one with the best evaluation measure.
6. Once the model is created (*k* chosen, etc.), use the same training set against unknown data you want to classify.