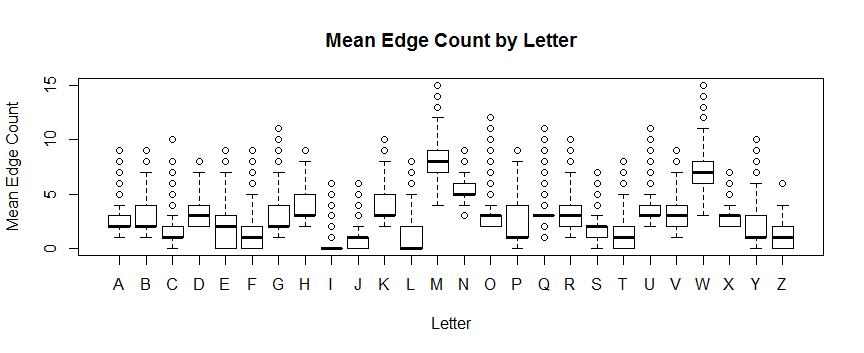
David Lowe

EDA #7

1. Digitizing documents has become more important as the public has sought access to genealogy, history, and literature. These important documents before the computer age (and some after) are contained in hand-written manuscripts that can be illegible for the public, and may be misinterpreted, or completely useless unless correctly transcribed to readable text. Digitizing documents is tedious and error prone when wrested completely on the ability of humans to read and copy. For this reason, we try to have computer predict letters and words by the characteristics of digital scans of the documents. If computers can take on most the load of successfully digitally transcribing manuscripts, humans need only proofread the digitized copies. This would greatly speed up the process of document digitization, and provide a massive wealth of documents to the public that is hidden from the public’s eye.
2. The data provided has 19,999 letters from the English alphabet, taken from a document, each having been “person-verified.” Each of these letters is also characterized with 16 attributes coming from the digitization of the document. Data such as pixel count, number of edges, position in the letter box, height and width of the letter box, etc. are recorded and will be used as predictors for the letters. Each of these predictor variables is either numeric or integer valued. The figure below shows the relationship with the mean number of edges in the letter, along with the letter it corresponds to. As can be seen, there is a wide range of number of edges, indicating that different letters can be distinguished by the number of edges in the letter. This, along with other variables should be useful in identifying letters.
3. To predict the letter, given the digitized characteristics, a classification method will need to be employed. I believe that random forests have an algorithm that groups similar data and can produce accurate predictions through this classification method. This would be a technique that I think could be effective in a setting such as this, where there are 26 classification groups, and several variables to explain them.
4. I have never used random forests to predict and I am not sure on the theory behind them. I also do not know how to include uncertainty when using a classification method like this.