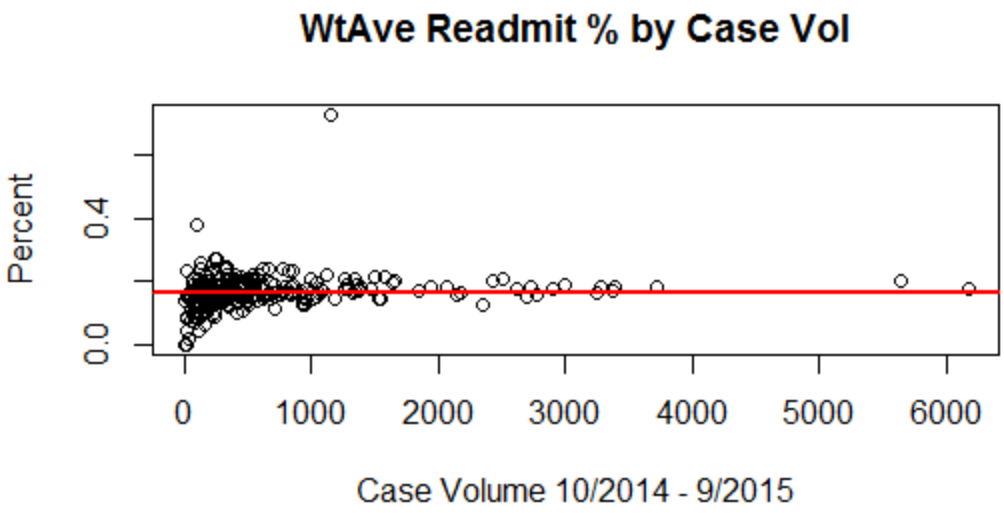
Fun With Outliers

One of the more overlooked things required for a good analytic product is detection and handling of outliers. Outliers have an annoying way of skewing our results, giving false results and causing uncomfortable questions when presenting our findings. For example, for the 30 day COPD readmit module for the Member Variation Analyzer, I pulled the 30 day COPD readmit percentages for all CCC Members. When plotting the Readmission percentage for each Member (which itself was an average of each Members hospital readmit percentage, weighted by case volume), I got the following plot:



The red horizontal line is the cohort mean of 17.8%. For the most part, this distribution follows the classic pattern where instances of low case volume show greater variability, but as the case volume increase, the data points tighten up around the population mean (to learn more about this phenomenon, read about the Law of Large Numbers).

However, there is an odd ball up around 70% readmission rate with a little over 10,000 encounters. The rest of the data points at that case volume are pretty tight around the mean. So, what's going on with this data point?

A bit of investigation found that this Member had an average severity of 1.3 compared to a cohort mean severity of 2.4. (Note: Severity is measured on a scale of 1 to 4 with 1 being very health and 4 being very unhealthy). So, a large volume Member with dramatically healthier patients than the cohort has a ridiculously high readmission percentage?

WHAAAAAAT?

At this point, I have to conclude that this particular Member just has bad data and I will exclude it from this analysis, which I did using the following code in R.

```
CohortWtAveReadmitPercent <- with(byMember, weighted.mean(WtAveReadmitPercentage, TotalEnc,
na.rm = TRUE))
CohortSdReadmits <- sd(byMember$WtAveReadmitPercentage, na.rm = TRUE)
byMember$z_score <- with(byMember, abs((WtAveReadmitPercentage -
CohortAveReadmits)/CohortSdReadmits))
byMember <- subset(byMember, z_score < 3)
```

The key point to keep in mind is that we should always ask questions about the data we are working with, particularly: "Is all the data in this data set suitable for analysis?" Thus, before undertaking any deep or complex analysis, we should first invest some time exploring the distributions of the key variables to to identify and remove problematic data points.