Data Log

1. Changed column names to reflect attributes: v.min, v.max, v.mean, v.mean distribution, v.contrast, v.entropy, v. second angular momentum, IR.Min, IR.Max, and IR.Mean

* Noticed that min and mean were switched, so changed column 1 and column 3, as well as column 8 and column 10

2. Change all ? to NA

3. Change all data types to numeric

4. Removed rows 483, 514, 549, 734, 793, 801, 924 since they have more than 20% of data missing

5. Removed outliers from attribute columns

* Visible Minimum greater than 80
* Visible Mean greater than 150 because there was a gap between 150 and the next data points above it
* Visible Contrast greater than 2500 because the histogram drops off after 2500 and there are a few instances in the box plot significantly further than the rest of the data (kept very small data points even though the range was large because there was such a large amount of instances less than 1)
* Visible Second Angular momentum removed outlier data point 888
* IR mean removed data less than 100 because although there are few points between 150 and 100, I wanted to keep as much data as possible
* IR Min removed data less than 40 because there is a small bump in the data after the relatively continuous distribution ends

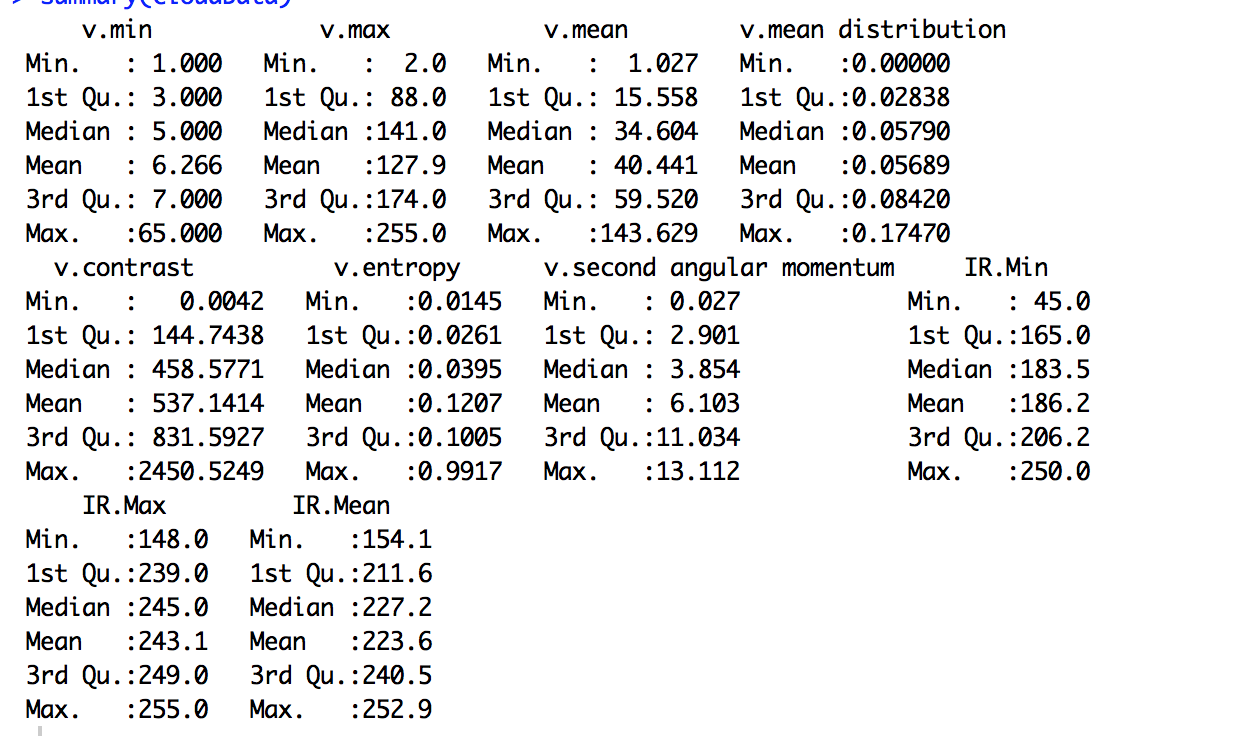
6. Use Correlation to impute values for which the correlation is greater than .9

* Replace the contrast value for rows 135, 158, 428, 647, 674, and 572 with -178.4 + 12579.4\*(visible mean distribution)
* Ignored correlation between visible mean and IR mean because they were NA at the same values

7. Use kNN imputation to estimate all other values since it takes into account the more subtle structures of the data

* Use n = 32 since = 31.82

8. Normalize data between 0 and 1

* Summary of data before in case we need to transform it back

Comments about the Histograms:

The histogram of second angular momentum is particularly interesting because it follows a bimodal distribution centered around 3 and 11. This shape implies that clouds’ second angular momentums are relatively normally distributed around two values. Visible Min, visible mean, visual entropy, and visual contrast are all strictly decreasing implying that lower values of these attributes are more common. This is an interesting contrast to Infrared Mean and Infrared Max, which is heavily left skewed. This shows that there are more instances with higher levels of these attributes. The histogram of Infrared Mean is also interesting because the frequency is almost strictly increasing, except for a sharp drop around 260. Similarly, the histogram for Mean distribution is relatively constant, and then decreases significantly after .1. The mode of Infrared Min looks like it is in the middle of the range, yet there are many more instances of values above the mode than below it. This implies that the mean and median of the data are above the mode.

Many of the attributes could benefit from a log transformation before scaling to minimize their skewness, such as Visible Max. I would avoid a log transformation on the bimodal distribution of second angular momentum because it might eliminate or smooth over important information about this distribution.