**Homework 1: DS 4559**

**Cleaning/Pre-processing Data**

1. Your final scrubbed and scaled dataset
2. Your complete, thorough data log, describing your steps (complete sentences are not necessary; this can just be a list)
3. Your annotated R code
4. Answer(s) to the following: Did you find any interesting histograms? Describe. What did their shape(s) imply? Do you think attributes in this dataset would benefit from a log transformation to make them more Gaussian before scaling? Explain.

2.

Install and library DMwR, ggplot2, and data.table

Import CloudData1

Insert columns from info file

Replace ? with NA

Transform all columns to numeric data

Summarize data

Switch min and max column names because upon looking at data, they are clearly switched

Look at boxplots for each attribute and determine outliers (annotated in code)

Create outlier replace function

Replace outliers in each attribute with NA

Remove rows with more than 20% of its data missing

Find correlations

Find linear trend between mean distribution and contrast and for IR mean and mean

Find which values are missing in each of these attributes

Replace values for each of these four attributes using correlation imputation

Use sum of na to find where missing values still exist

Normalize data by creating norm function

Apply norm function to each attribute

Use knnImputation to determine the rest of the missing values (error with mean imputation)

Import data back into file

4

The min is understandably skewed towards the lower end and the max towards the upper end, but the mean is also skewed towards the lower end and could use a log transform to appear more standard before normalizing. The 2nd angular momentum had 2 peaks which was surprising to me, but lent itself well to knnImputation rather than mean imputation. IR min, max, and mean did not follow the same trends as min, max, and mean, which was surprising. It is possible that the absence of one leads to a greater presence of the IR counterpart, and it would be helpful to discuss with a domain expert to verify.