Charlotte Searle DS Homework #1

Cloud Data Log:

- Read in the csv file
- Changed attribute names according to the info file
- Read values in as numbers instead of characters
- Switched VIS_mean and VIS_min, as well IR_mean and IR_min attribute names because the data made it clear they were wrongfully named
- Outlier Detection
  - Defined outlierReplace function
  - VIS_min: replaced outliers with values greater than 80
  - VIS_max: looks good (no unreasonable values)
  - VIS_mean: replaced outliers above 140
  - VIS_mean_distribution: looks good
  - VIS_contrast: replaced outliers above 2750
  - VIS_entropy: looks good (no unreasonable values)
  - VIS_second_angular_momentum: replaced outliers above 200
  - IR_min: replaced outliers less than 25
  - IR_max: looks good (no unreasonable values)
  - IR_mean: replaced outliers below 125
- Found a correlation between VIS_contrast and VIS_mean_distribution
  - Found all lines where VIS_contrast was NA
    - Lines: 135, 158, 569, 644, 671
  - Created a linear model using VIS_contrast and VIS_mean_distribution
    - Coefficients: Intercept = -181, VIS_mean_distribution = 12,646
  - Assigned the missing values of VIS_contrast using correlation imputation based on the equation above found from making the linear model
- Used kNN imputation for the rest of the data, with k=sqrt(1008)
- Created a normalizing function with help from the following site: http://vitalflux.com/data-science-scale-normalize-numeric-data-using-r/
- Note: the following are the max/mins of each attribute before using normalizing function
  - VIS_min
    - min: 1.000
    - max: 65.000
  - VIS_max
    - min: 2.0
    - max: 255.0
  - VIS_mean
    - min: 1.027
    - max: 136.074
  - VIS_mean_distribution
    - min: 0
    - max: 0.17470
  - VIS_contrast
    - min: 0.0042
    - max: 2519.9666
  - VIS_entropy
    - min: 0.0145
    - max: 0.9917
  - VIS_second_angular_momentum
    - min: 0.027

    - max: 13.112
  - IR_min
    - min: 45.0
    - max: 250.0
  - IR_max
    - min: 148.0
    - max: 255.0
  - IR_mean
    - min: 130.7
    - max: 252.9
- Normalized the entire dataset using the above function
- Saved the clean datafile, “clouddata_clean.csv”

Answer to Homework Questions:

Left-skew: VIS_min, VIS_mean, VIS_mean_distribution, VIS_contrast, VIS_entropy
slight right-skew: VIS_max
Right-skew: IR_max, IR_mean, IR_min
Bi-modal: VIS_second_angular_momentum around 0-4 and 7-10

I found that almost all of the attributes which were attained through visible satellite imagery are left skewed. The fact that the values are left skewed suggests that we would benefit from logarithmically transforming such attributes in order to make them more Gaussian. On the other hand, VIS_second_angular_momentum is bimodal, with Gaussian-like curves around values of 0-4 and around values of 7-10. Hence, we would not want to use a log transform function because it could smooth over critical information. Namely, that there are two Gaussian-like groups associated with second_angular_momentum. Finally, all attributes collected via infrared imagery (IR) are right-skewed, and hence would also benefit from a log transformation.