1. Descriptive analysis of additives:
   1. We will use the Pearson correlation coefficient for all columns. Assuming the below:
      1. Correlation coefficients of greater than 0.5 are high
      2. Correlation coefficients of 0.3-0.49 are moderate
      3. Correlation coefficients below 0.3 are weak

We end up with the below set of correlations:

**High negative correlation between a and e of -0.5420521997085237**

**High positive correlation between a and g of 0.8104026963400857**

Medium negative correlation between a and d

Medium negative correlation between c and d

Medium positive correlation between d and f

Medium negative correlation between c and g

Medium negative correlation between f and g

Medium positive correlation between b and h

Medium negative correlation between c and h

Medium positive correlation between d and h

All the other correlations are low. Additives *a* and *e* are likely to be antagonistic to each other, while additives *a* and *g* are likely to be required in tandem for the proper functioning of most petrol mixes.

* 1. Next, we test for normalized standard deviation (also known as the coefficient of variance, which is given by standard deviation/mean) for each additive, and rank them in descending order:

**h: 2.840494**

**i: 1.709171**

**f: 1.312109**

c**:** 0.537303

d**:** 0.345538

g**:** 0.158888

b**:** 0.060905

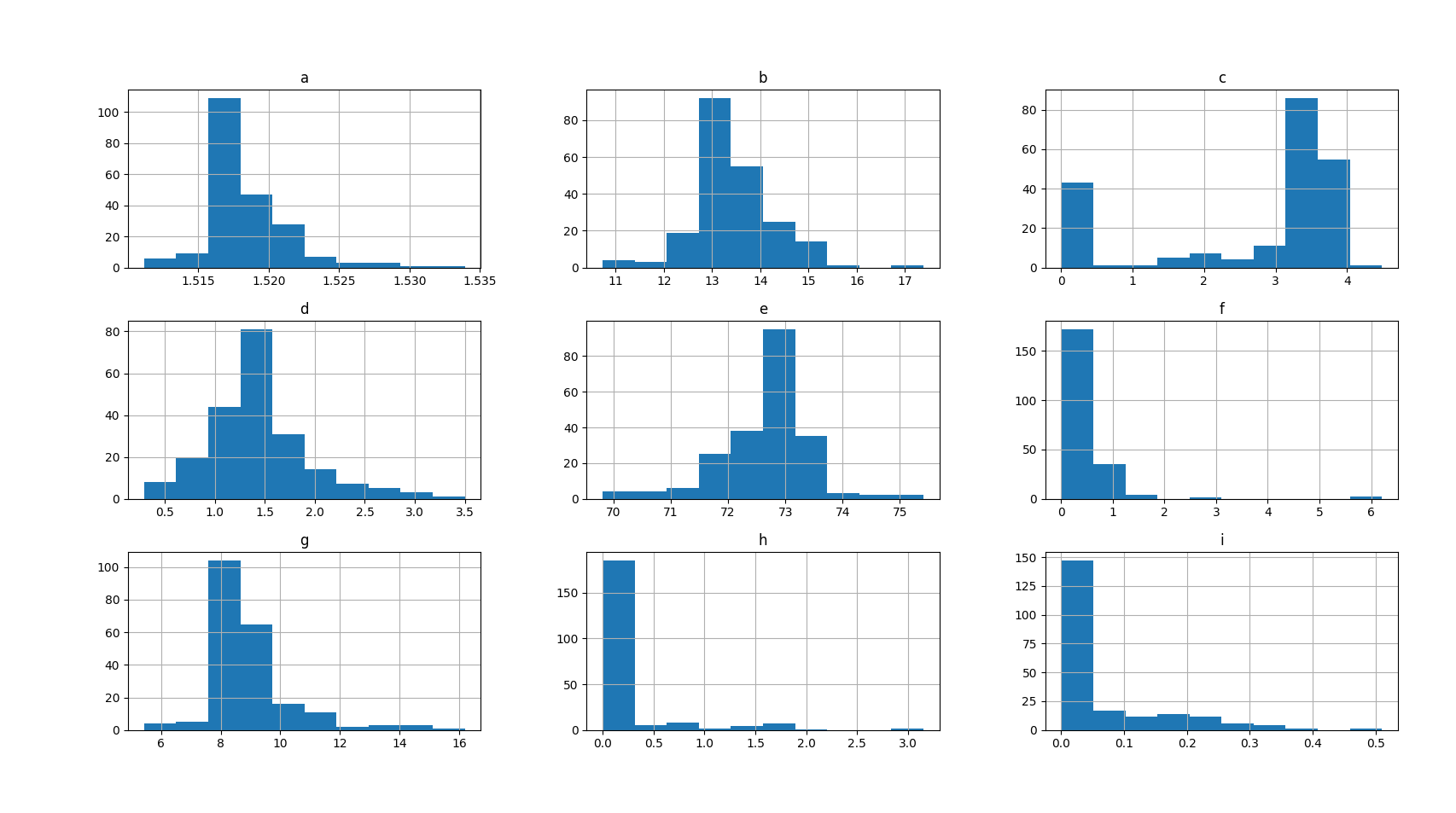
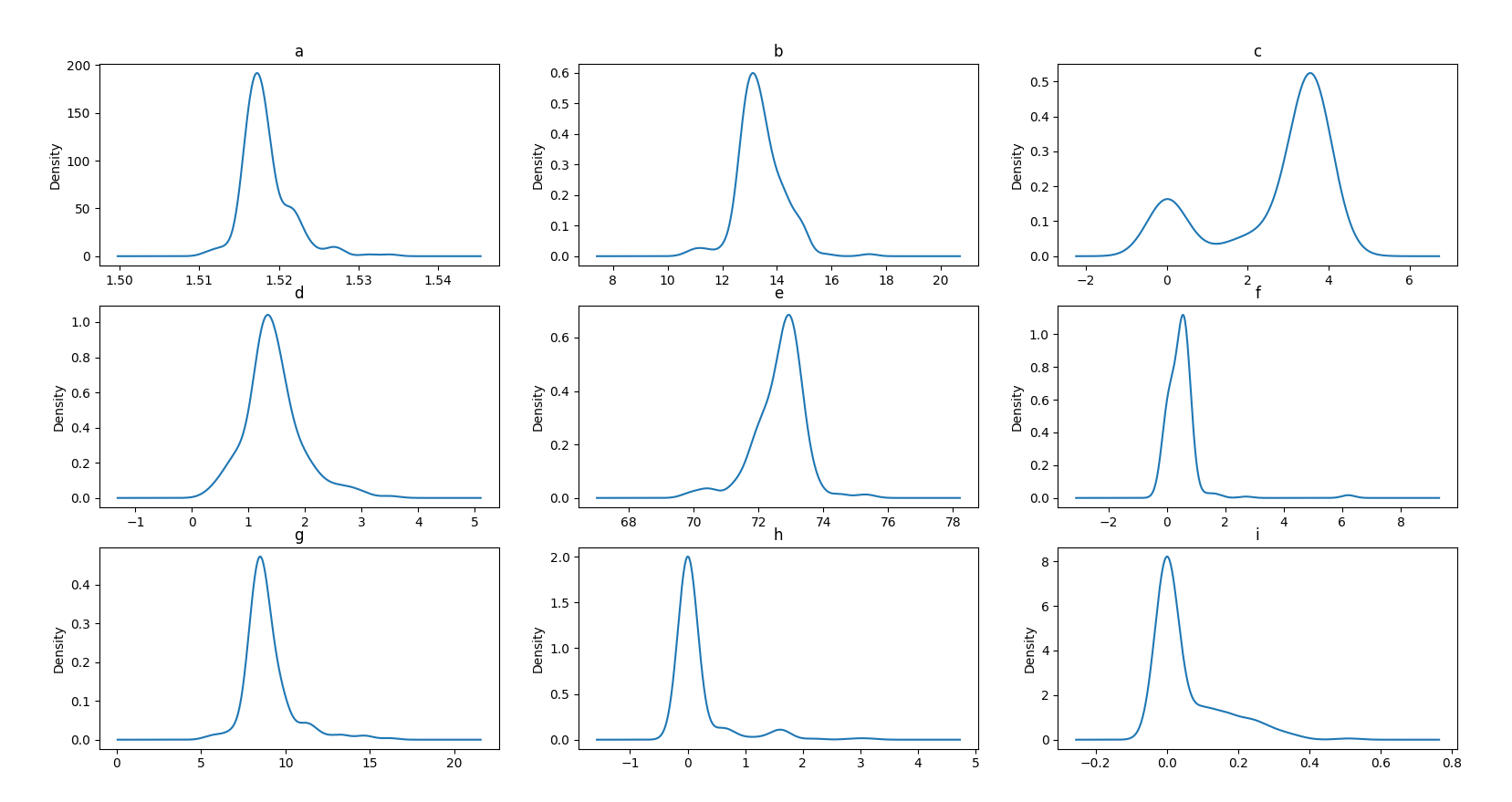
e**:** 0.010661

a**:** 0.002000

Additives *h*, *i*, and *f* are ingredients with high variability, as they have high coefficients of variance (greater or equals to 1). We would reasonably expect these additives to be determining factors when it comes to clustering the mixes.

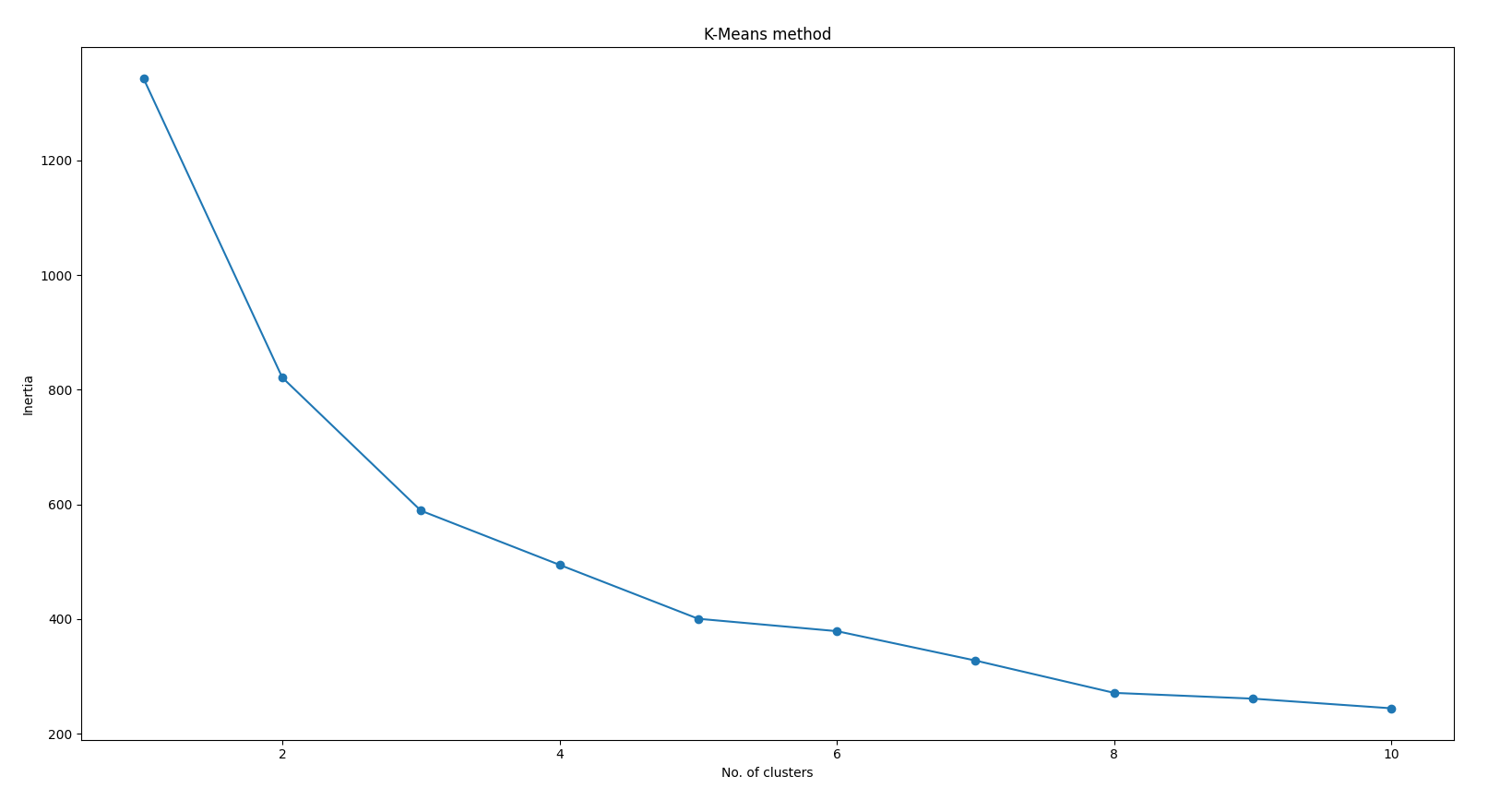
1. Graphical analysis of additives

For a graphical analysis, a histogram for each additive should be sufficient to illustrate the usual concentrations that they are added in:



We can see that all the additives are unimodal, and only *c* is bimodal. This strongly suggests that *c* plays a major role in more than one formulation at the very least.

1. I have chosen to use k-means clustering for this:



The elbow method suggests that there should be 3 clusters in the data.

Since K-Means clustering is a linear method, we should also verify our results with a nonlinear method (OPTICS, a variant of DBScan). The labels returned were:

[-1 0 1 2]

Since noisy labels were given a value of -1, this means that OPOTICS predicted there would be 3 clusters, in line with our KNN prediction.