**DATA ANALYSIS**

In Sum(+/-%,ASK)-name graph it can be seen that Natural Gas 1st has the highest +/- % value fluctuation which is -6.870 and Lead 1st has the least +/- % fluctuation value which is -0.030.

The total ASK value is the highest for Silver 1st which is 125,052,2nd highest ASK is GOLD 1ST which is 51347 and 3rd highest ASK is 30645.

K-means procedure splits the data into K segments.  Each segment has a centroid that corresponds to the mean value for the members in that segment. The objective of the algorithm is to place the centroids such that the total of the sum of distances between centroids and members in respective segments is as small as possible.

**Analysis of Variance:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | **Model** | |  | **Error** | |
| **Variable** |  | **F-statistic** |  | **p-value** |  | **Sum of Squares** | **DF** |  | **Sum of Squares** | **DF** |
| **Sum of Ask** |  | 6.69 |  | 0.0006551 |  | 1.08 | 4 |  | 1.13 | 28 |
| **Sum of +/- %** |  | 5.697 |  | 0.001745 |  | 1.244 | 4 |  | 1.529 | 28 |
| **Inputs for Clustering**   |  |  | | --- | --- | | **Variables:** | Sum of +/- % | |  | Sum of Ask | | **Level of Detail:** | Name | | **Scaling:** | Normalized | |  |  |   **Summary Diagnostics**   |  |  | | --- | --- | | **Number of Clusters:** | 5 | | **Number of Points:** | 33 | | **Between-group Sum of Squares:** | 2.3243 | | **Within-group Sum of Squares:** | 0.33474 | | **Total Sum of Squares:** | 2.6591 | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  | **Centers** | | | | | | | |  | | **Clusters** |  | **Number of Items** |  | **Sum of +/- %** | | | | **Sum of Ask** | | | |  | | **Cluster 1** |  | 9 |  | -2.2611 | | | | 2927.5 | | | |  | | **Cluster 2** |  | 1 |  | 0.68 | | | | 1.2505e+05 | | | |  | | **Cluster 3** |  | 18 |  | 0.11833 | | | | 2229.2 | | | |  | | **Cluster 4** |  | 2 |  | 1.045 | | | | 40996.0 | | | |  | | **Cluster 5** |  | 3 |  | -5.2167 | | | | 212.35 | | | |  | | **Not Clustered** |  | 0 |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |  |  |  |

Total sum of squares represents a measure of how a dataset varies around a central number(like the mean):2.6591

Sum of Squares total=Sum of squares within+Sum of squares between(SST=SSW+SSB)

Sum of squares within represents variation due to individual differences in the score.

Error Sum of squares(SSE) is the sum of the squared differences between each observation and it’s group mean.It can be used as a measure of variation within a cluster .If all cases within a cluster are identical the Sum of squares error would then be equal to 0.

### **Describe Clusters – Summary Tab**

The Summary tab identifies the inputs that were used to generate the clusters and provides some statistics that characterize the clusters.

#### **Inputs for Clustering**

**Variables**

Identifies the fields Tableau uses to compute clusters. These are the fields listed in the Variables box in the Clusters dialog box.

**Level of Detail**

Identifies the fields that are contributing to the view’s level of detail—that is, the fields that determine the level of aggregation. For details, see [How dimensions affect the level of detail in the view](https://help.tableau.com/current/public/desktop/en-us/datafields_typesandroles.htm#DimLOD).

**Scaling**

Identifies the scaling method used for pre-processing. Normalized is currently the only scaling method Tableau uses. The formula for this method, also known as min-max normalization, is (x – min(x))/(max(x) - min(x)).

#### **Summary Diagnostics**

**Number of Clusters**

The number of individual clusters in the clustering.Which is 5 here.

**Number of Points**

The number of marks in the view.Which is 33 here.

**Between-group sum of squares**

A metric quantifying the separation between clusters as a sum of squared distances between each cluster’s center (average value), weighted by the number of data points assigned to the cluster, and the center of the data set. **The larger the value, the better the separation between clusters.**

**Within-group sum of squares**

A metric quantifying the cohesion of clusters as a sum of squared distances between the center of each cluster and the individual marks in the cluster. **The smaller the value, the more cohesive the clusters.**

**Total sum of squares**

Totals the between-group sum of squares and the within-group sum of squares. The ratio (between-group sum of squares)/(total sum of squares) gives the proportion of variance explained by the model. Values are between 0 and 1; **larger values typically indicate a better model.** However, you can increase this ratio just by increasing the number of clusters, so it could be misleading if you compare a five-cluster model with a three-cluster model using just this value.

#### **Cluster Statistics**

For each cluster in the clustering, the following information is provided.

**# Items**

The number of marks within the cluster.

**Centers**

The average value within each cluster (shown for numeric items).

**Most Common**

The most common value within each cluster (only shown for categorical items).

### **Describe Clusters – Models Tab**

Analysis of variance (ANOVA) is a collection of statistical models and associated procedures useful for analyzing variation within and between observations that have been partitioned into groups or clusters. In this case, analysis of variance is computed per variable, and **the resulting analysis of variance table can be used to determine which variables are most effective for distinguishing clusters.**

Relevant analysis of variance statistics for clustering include:

#### **F-statistic**

The F-statistic for one-way, or single-factor, ANOVA is the fraction of variance explained by a variable. It is the ratio of the between-group variance to the total variance.

**The larger the F-statistic, the better the corresponding variable is distinguishing between clusters.**

#### **p-value**

The p-value is the probability that the F-distribution of all possible values of the F-statistic takes on a value greater than the actual F-statistic for a variable. If the p-value falls below a specified significance level, then the null hypothesis (**that the individual elements of the variable are random samples from a single population**) can be rejected. The degrees of freedom for this F- distribution are (k - 1, N - k), where k is the number of clusters and N is the number of items (rows) clustered.

**The lower the p-value, the more the expected values of the elements of the corresponding variable differ among clusters.**

#### **Model Sum of Squares and Degrees of Freedom**

The Model Sum of Squares is the ratio of the between-group sum of squares to the model degrees of freedom. The between group sum of squares is a measure of the variation between cluster means. If the cluster means are close to each other (and therefore close to the overall mean), this value will be small. The model has k-1 degrees of freedom, where k is the number of clusters.

#### **Error Sum of Squares and Degrees of Freedom**

The Error Sum of Squares is the ratio of within-group sum of squares to the error degrees of freedom. The within-group sum-of-squares measures the variation between observations within each cluster. The error has N-k degrees of freedom, where N is the total number of observations (rows) clustered and k is the number of clusters.

The Error Sum of Squares can be thought of as the overall Mean Square Error, assuming that each cluster center represents the "truth" for each cluster.





















