**The Task:**

The purpose of this analysis and case study is Segmentation: Identification and profiling of customer groups by key business parameters price paid, sales, discounts, units sold of a software company.. A statistical model is then adopted to further the analysis and arrive at the results and interpretation.

**The Dataset:**

The dataset contains the following variables:

****

**The statistical model:**

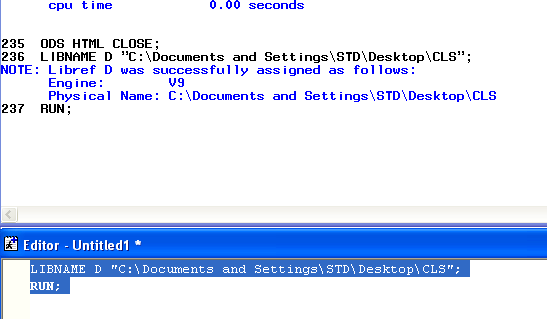
We have adopted the clustering analysis in this case. **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory [data mining](https://en.wikipedia.org/wiki/Data_mining), and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [machine learning](https://en.wikipedia.org/wiki/Machine_learning), [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [image analysis](https://en.wikipedia.org/wiki/Image_analysis), [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics), [data compression](https://en.wikipedia.org/wiki/Data_compression), and [computer graphics](https://en.wikipedia.org/wiki/Computer_graphics).

There are two types of clustering.

* [***Hierarchical clustering***](https://en.wikipedia.org/wiki/Hierarchical_clustering): Connectivity-based clustering, also known as [*hierarchical clustering*](https://en.wikipedia.org/wiki/Hierarchical_clustering), is based on the core idea of objects being more related to nearby objects than to objects farther away.
* [***K-Means clustering***](https://en.wikipedia.org/wiki/K-means_clustering): In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to *k*, [*k*-means clustering](https://en.wikipedia.org/wiki/K-means_clustering) gives a formal definition as an optimization problem: find the *k* cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

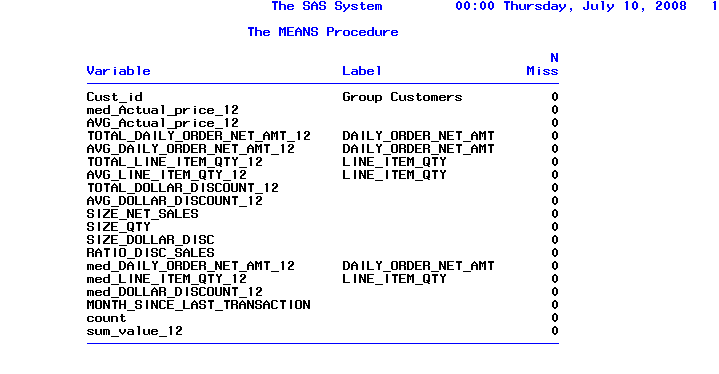
In this case we have used K-Means clustering method.

**Setting up the SAS model by loading the required libraries:**



**Data:**

In the next step, the data is read into the SAS environment from the file and checked for any missing values.



Luckily, it didn’t contain any missing value.

**Factor Analysis:**

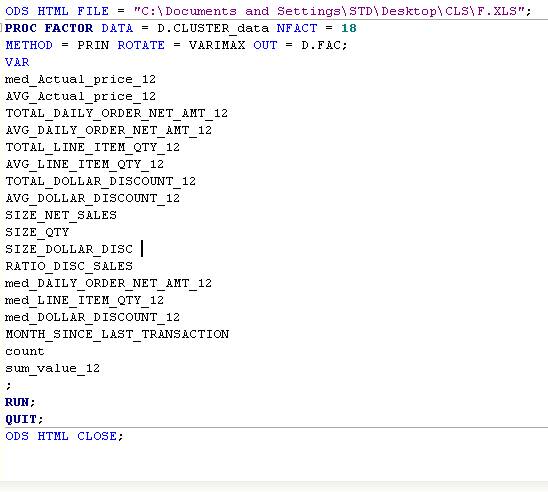
Reduce number of variables and lower multicollinearity. Factors are linear combination of the variables.

NFACT: NO OF FACTORS GENERALLY KEPT SAME AS THE NUMBER OF VARIABLES

METHOD: PRIN (SHORT FORM FOR PRINCIPAL COMPONENT ANALYSIS)

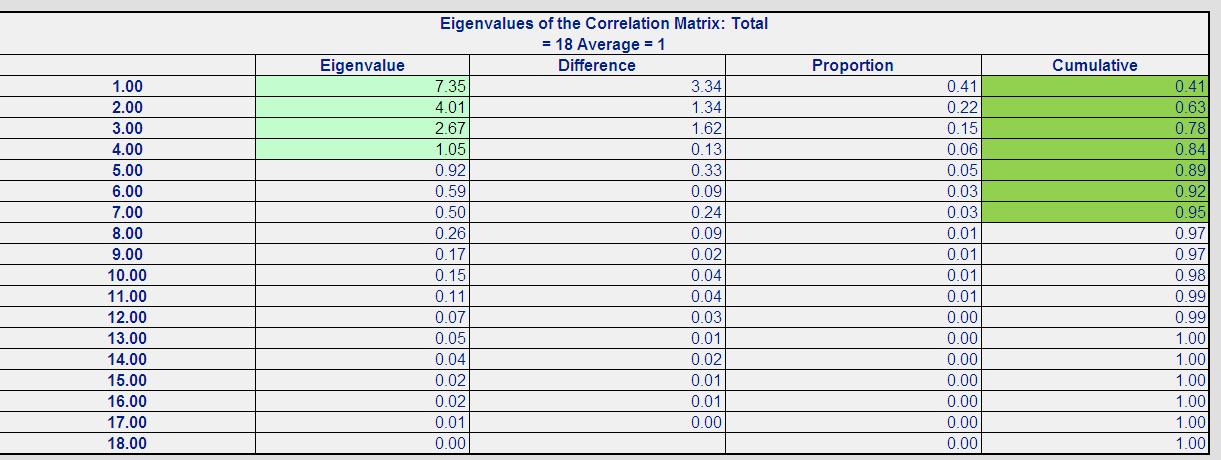
ROTATE : VARIMAX (ROTATION METHOD ENSURING ORTHOGANILITY OF FACTORS)

OUT : OUTPUT FILE WHICH STORE THE FACTOR SCORES;



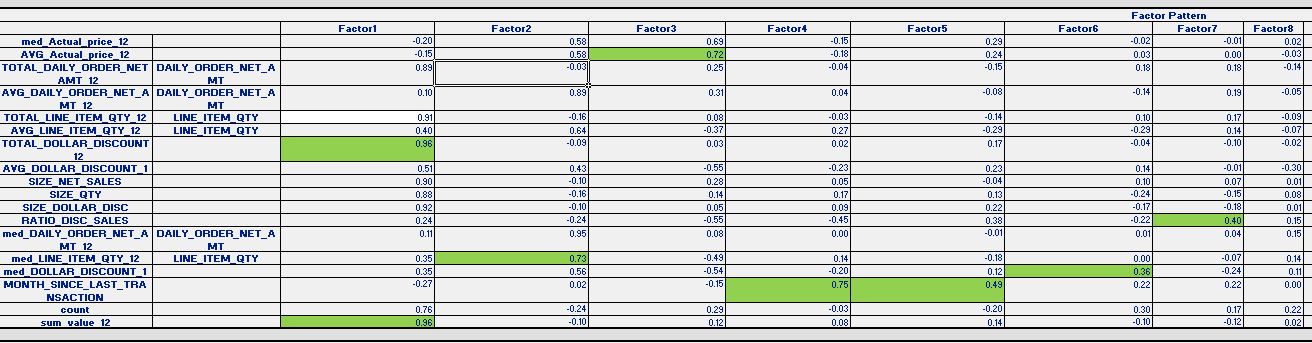
A total of 18 factors we have.

Next we will look into the factors that give eigenvalue of more than 1. But we will select those factors whose cumulative frequency is at most 95%.



In this case we will select first 7 factors.

Consider those factors which account for 90-95% of the total variation in the data. For these selected factors we consider the variables with high loadings. This variable set is used for further analysis.

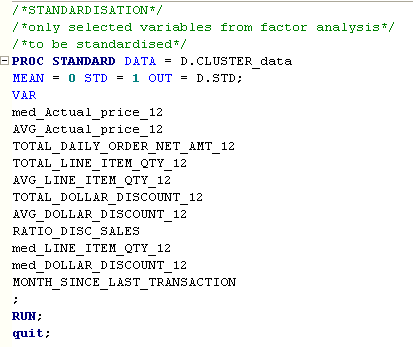


126 Variables – 7 Factors which account for 90-95% variation.

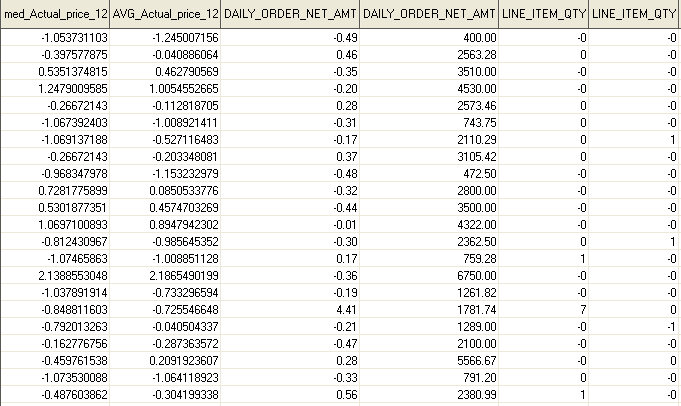
– 7 Variables with high loading (will have lower multicollinearity than the original data) are chosen from these 7 factors.

**Standardization of the data:**

Database normalization/standardization is the process of structuring a relational database in accordance with a series of so-called normal forms in order to reduce data redundancy and improve data integrity.



We must always normalize the data set so that the output remains unbiased.



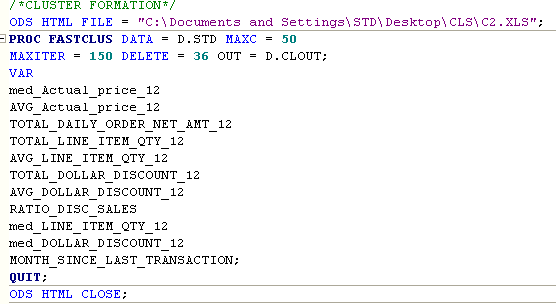
**Cluster Formation:**

With the standardized cluster formation is done. K – Means Clustering is done using

PROC FASTCLUS DATA = D.STD MAXC = 50 (maximum number of clusters to start with)

MAXITER = 150 (maximum Number of Iterations) DELETE = 36 (Minimum Observation per Cluster – generally kept at 5% of total observation where total observations = 728) OUT = D.CLOUT (saves the cluster membership variable);

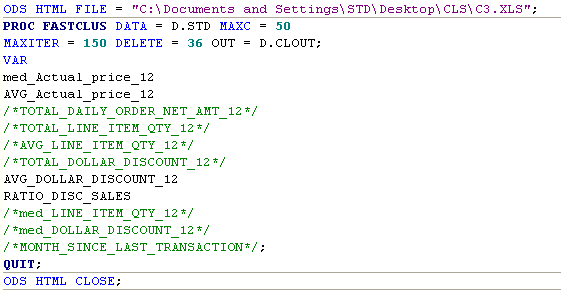
After the 1st cut we alter the variable set till the following checks of optimality for cluster are fulfilled.



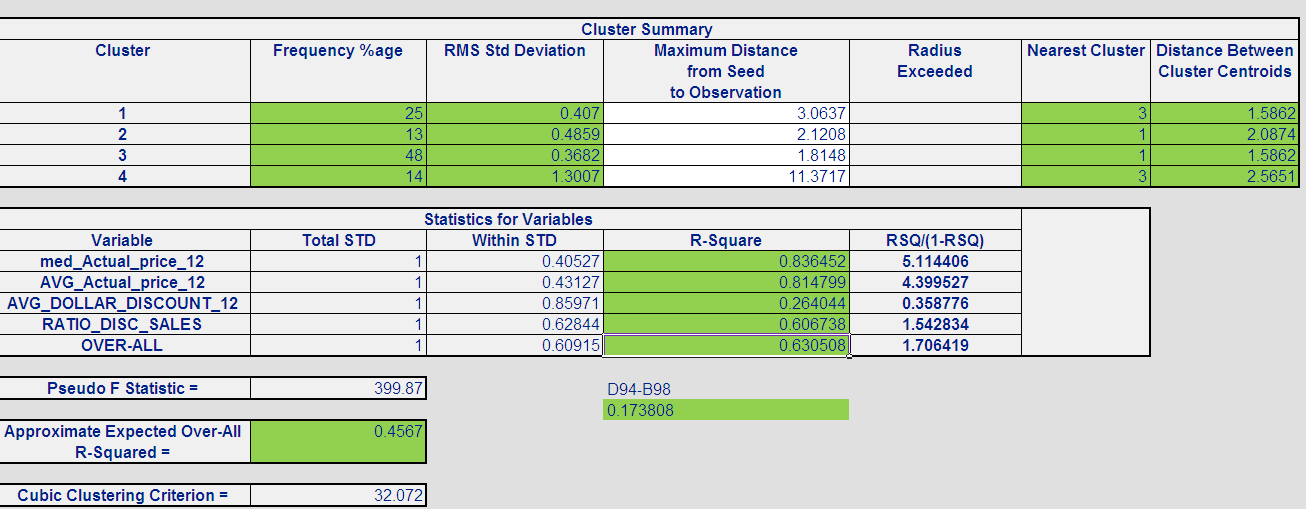
**Checks for cluster:**

1. **Individual R-Squared** >= 0.25. Every variable used in cluster formation generates an R-Squared. This measures the worth of the variable in the cluster formation. The final model must have only those variables for which R-Squared >= 0.25
2. **Overall R-Squared** >= 0.5. This measure the overall goodness of fit of the model and should be >= 0.5.
3. **Approximate Expected Overall R-Squared** > =0.3 (This is the R-Squared which the model would have generated if there was no Multicollinearity among the variables used in cluster formation). It should be >= 0.3.
4. **The Difference Between Overall R-Squared and Approximate Expected Overall R-Squared** Should Not Be Greater Than 0.2. A higher difference indicates unacceptable amount of Multicollinearity among the variables used in cluster formation.
5. **RMS Standard Dev** < =1.4. This is a measure of within cluster homogeneity. It should be < =1.4 for each cluster. A higher value for any cluster indicates presence of outliers in that cluster.
6. **Distance between Cluster Centroids** >= 1.4. This is a measure of across cluster heterogeneity. The distance between centroid of any cluster with that of the nearest cluster should be >= 1.4
7. **Number of Clusters** Should Be Between 4 and 15.
8. **Percentage of Frequency** in Each Cluster Should Be < = 35.

**Final Model:**



Here those variables which are insignificant we made them statement(green) and ran the codes so that we can show the changes we made.



Here all the checks have been fulfilled.