DATA VISUALIZATION AND MINING **STAT:4660**

PROJECT: SMART SUPPLY CHAIN

Dataset used for learning, examination and visualization using R

ALIYU BELLO

UNIVERSITY of Prince Edward ISLAND CAIRO Campus | Hosted by UNIVERSITIES OF CANADA IN EGYPT

CONTENT:

- - Overview.
- - Data Features.
- - Histogram for quantitative variables.
- - Bar chart for every qualitative variable.
- - Hoteling T2 test
- - Discriminant Analysis.
- - Clustering.
- - Principal Component analysis (PCA).
- - Multidimensional Scaling (MDS).
- - Canonical Correlation.
- - Outlier Detection.
- - Measures of Association.

SUPPLY CHAIN:

A supply chain is a system of all the people, businesses, resources, tasks, and technological advancements involved in producing and distributing a good. An entire supply chain is included, from the distribution of raw materials from the supplier to the producer to the final delivery to the customer.

Areas of important registered activities: Provisioning, Production, Sales, Commercial Distribution. It also allows the correlation of Structured Data with Unstructured Data for knowledge generation.

Categories: Data Mining, Supply Chain Management, Machine Learning, Big Data Analytics, Data visualization

STAGES OF SUPPLY CHAIN:

- **INTEGRATION:** Integration is crucial across your communications, information sharing, data analysis, and storage processes and begins at the strategic planning stage.

- **OPERATIONS:** In order to track output and predict production and distribution patterns, your operations need an accurate, real-time depiction of your inventory and production schedules.

- **PURCHASING:** When it comes to acquiring products for your supply chain and ensuring that you are benefiting from the most affordable prices and the most dependable products, the correct supply chain software does a lot.

- **DISTRIBUTION:** A part of your supply chain that can constantly be streamlined, improved, and corrected for better customer service and lower operational costs is the transport, delivery, and return of goods.

DATA FEATURES:

Finding inconsistencies, patterns, and correlations within significant data sets in order to identify outcomes is called data mining. Information could be utilized to lower risks, improve customer connections, raise profits, and more by implementing various techniques.

The data consist of 17 variables and 603 observations. It had a mass volume which was cut down to 15 quantitative variables and 2 qualitative variables in order to perform the analysis. The CSV centers around primary data set including.

Type: Cash, Debit, Payment, Transfer
Product Price
Order Profit Per Order
Customer Segment
Days for shipping
Days for shipment
Benefit per order
Sales per customer
Latitude
Longitude
Order Item Discount
Order Item Product Price
Order Item Profit Ratio
Order Item Quantity
Sales
Order Item Total

Type	Customer.Segment	Days.for.shipping..real.	Days.for.shipment..scheduled.
FALSE	FALSE	TRUE	TRUE
Benefit.per.order	**Sales.per.customer**	**Latitude**	**Longitude**
TRUE	TRUE	TRUE	TRUE
Order.Item.Discount	**Order.Item.Discount.Rate**	**Order.Item.Product.Price**	**Order.Item.Profit.Ratio**
TRUE	TRUE	TRUE	TRUE
Order.Item.Quantity	**Sales**	**Order.Item.Total**	**Order.Profit.Per.Order**
TRUE	TRUE	TRUE	TRUE
Product.Price			
TRUE			

For the analysis, a Data set of Supply Chains from the business DataCo Global was employed. Dataset for the supply chain that supports R software and machine learning algorithms.

HISTOGRAMS:

From the R output, below is a description from the visualization of the quantitative and qualitative variables of the data using histograms and bar plots.

Type: Bar plot Showing the value of quantity of each observation where Debit and Transfer are of higher quantity to Cash and Payment.

Product Price: Histogram showing the observation skewing to the right.

Order Profit Per Order: Histogram showing a normal distribution.

Customer Segment: Bar plot with higher number with Consumer and least with Home office.

Days for shipping: Histogram illustrating a leveled pattern of observations

Days for shipment: Histogram skewing to the left

Benefit per order: Histogram illustrating a normal distribution pattern.

Sales per customer: Histogram showing a pattern skewing to the right

Latitude: Histogram showing more of the observations between 15 and 20

Longitude: Histogram illustrating a skewed to the left pattern

Order Item Discount: Histogram illustrating a skewed to the right pattern with more observation on 0

Order item Discount Rate: The plot indicating a fluctuation on the pattern.

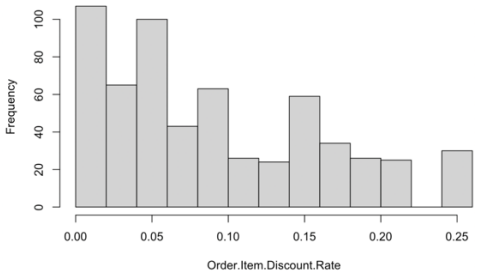
Order Item Product Price: Histogram showing iterative pattern between 0 and 500

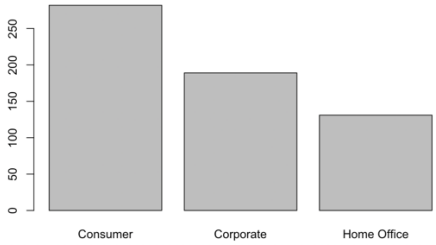
Order Item Profit Ratio: Histogram showing an extreme negatively skewed pattern with more observations between 0 and 5.

Order Item Quantity: Histogram showing a simple pattern approaching the right side

Sales: Histogram illustrating a fluctuation on the pattern with most observations between 0 and 500.

Order Item Total: Histogram showing normal pattern between 0 and 500 and few observations between 900 to 1400.





DISCRIMINANAT ANALYSIS:

This is a procedure that is done to analyze a research data when the dependent variable is categorical or qualitative, and the predictor or the independent variable is of an interval type. The quantity of categories the dependent variable possesses serves as a measure of discriminant analysis.

From the Supply Chain Dataset, below is the analysis using R output.

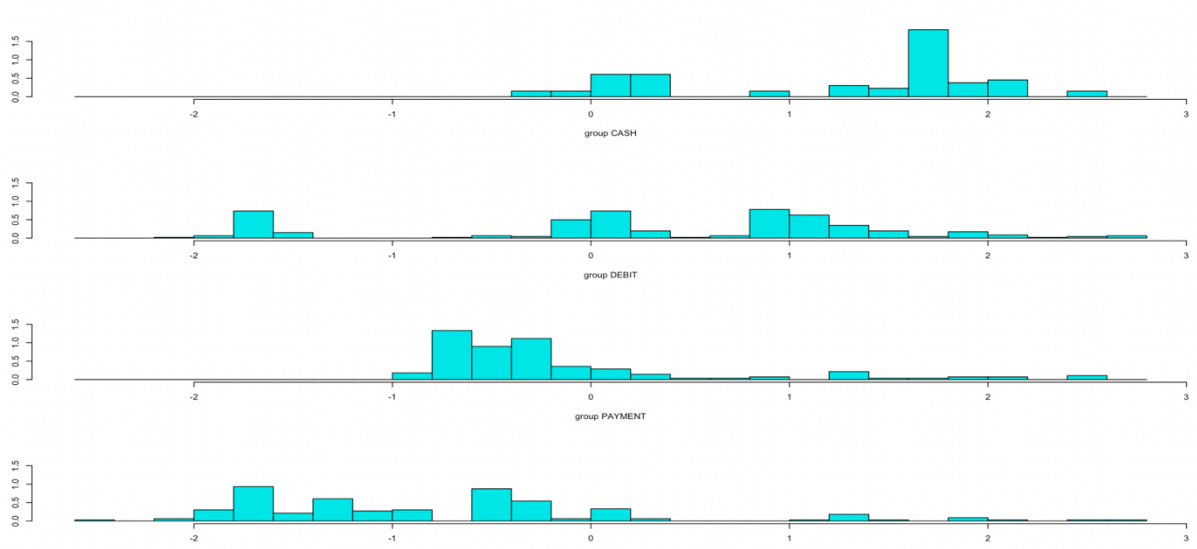
```
Call:
lda(Type ~ ., data = SupplyC)

Prior probabilities of groups:
     CASH     DEBIT   PAYMENT  TRANSFER
0.1096346 0.3837209 0.2308970 0.2757475
```

Above shows the first calling calculates prior probabilities of the groups indicating 10% of the payments goes to Cash, approximately 38% are of Debit, 23% of payments, and about 27% are of transfer.

```
Proportion of trace:
   LD1    LD2    LD3
0.8027 0.1363 0.0610
```

Above shows the coefficients of the linear discriminant for each of the variables. From the proportion separation of trace, we can see that LDA 1 whopping 80% achieved by the discriminant function. While, the second LDA got about 13%, and the third with a very small amount of approximately 1%.

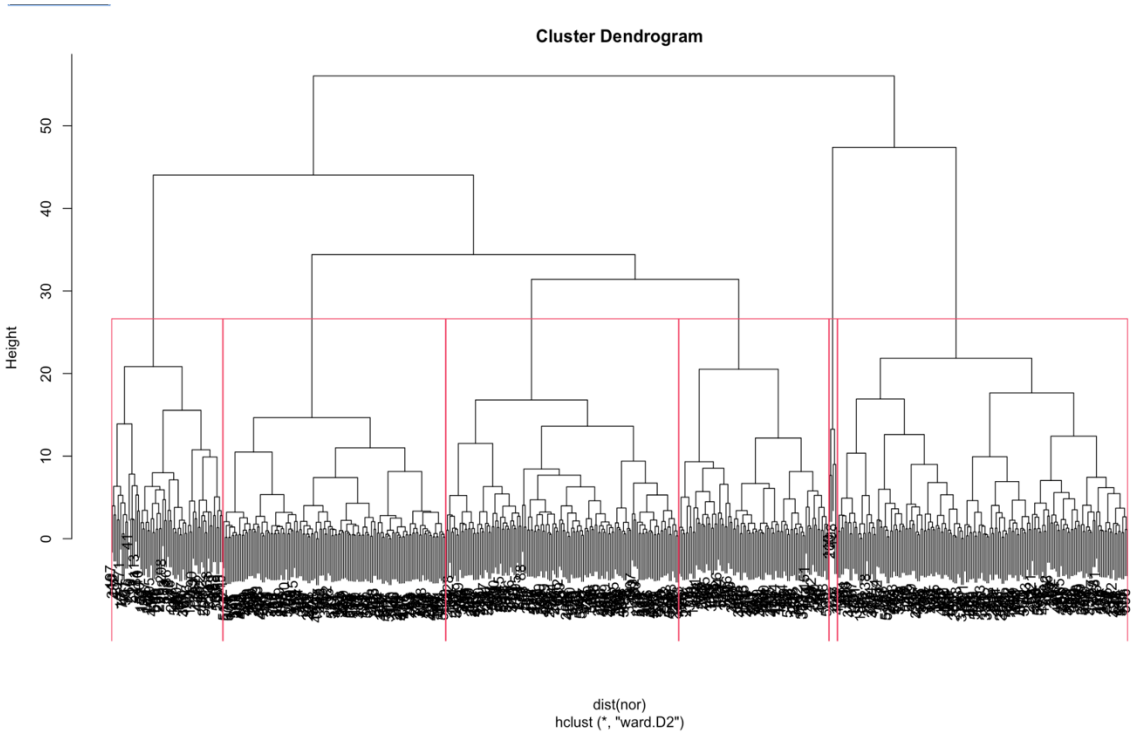


CLUSTERING

A group of abstract items can be organized into classes of related objects using the clustering technique. A method of dividing a set of data or objects into a number of important subclasses known as clusters is termed clustering**.** In this analysis two clustering techniques would be used to perform the method **Hierarchical** clustering and **K-means** clustering.

HEIRARCHIAL CLUSTERING:

An algorithm called hierarchical clustering, commonly referred to as hierarchical cluster analysis, divides objects into clusters based on how similar they are. The result is a collection of clusters, each of which differs from the others while having things that are generally similar.

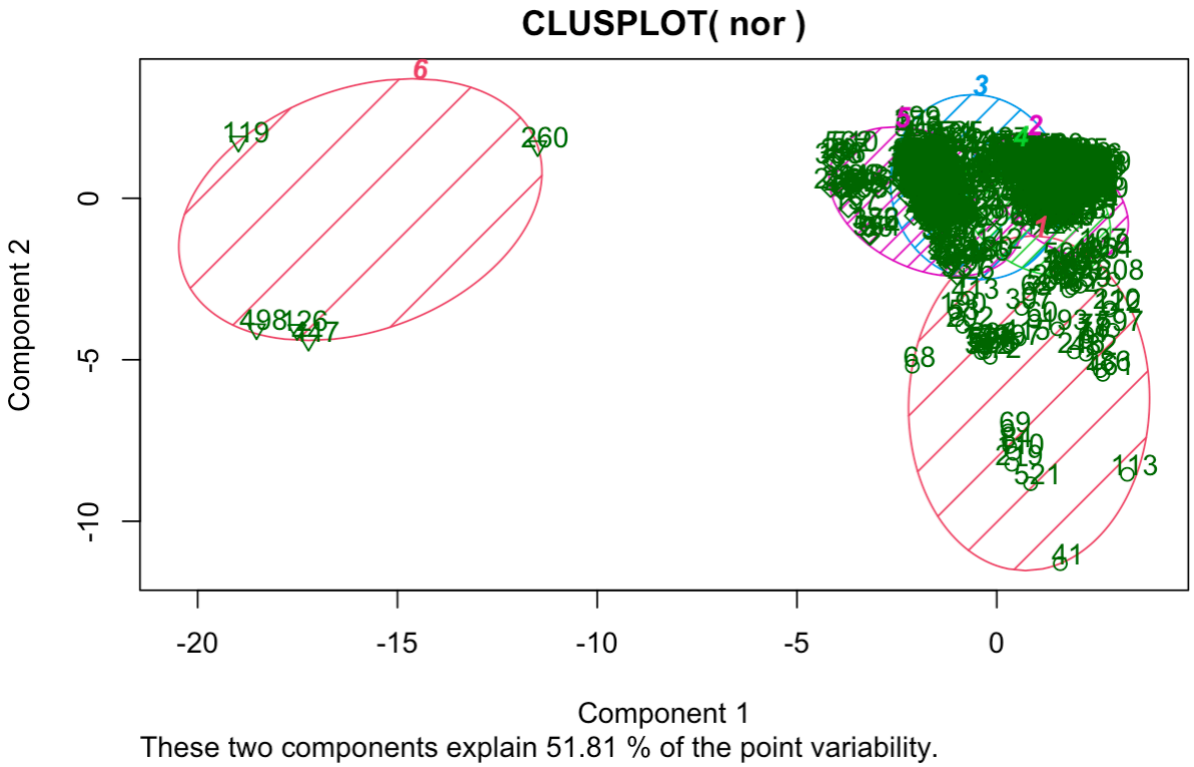


Above is a dendrogram with the from the hierarchical clustering technique indicating the within and between distances of the normalized data. The number of clusters chosen that describes the technique is **6**, while using the **Ward's** method. A red boarder line was used for better visualization of the dendrogram.

K-MEANS CLUSTERING:

K-means clustering aims to create groups out of comparable types of items. It determines whether two objects are similar to one another and clusters them.

Data mining's K-means technique uses an initial set of randomly chosen centroids as the starting points for each cluster to process the learning data, and iterative (repetitive) calculations are then used to optimize the positions of the centroids.



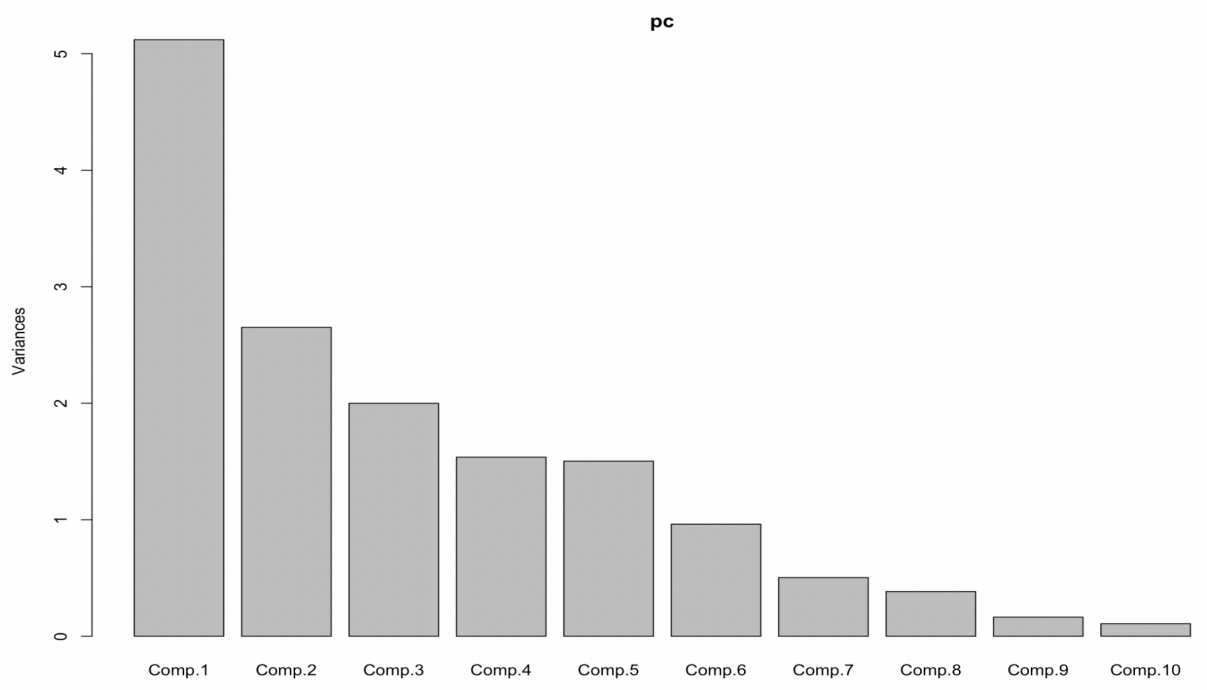

```
Within cluster sum of squares by cluster:
[1]  622.1630  474.9032  784.6052  684.9352 1352.2389  158.8019
 (between_SS / total_SS =  54.8 %)
```

Above is a clust plot from the K-means clustering technique showing the within and between sum of squares and the total sum of squares of the normalized data. This result to approximately **55%** which is decent for this data interpretation a result of the its volume. The number of clusters chosen that describes the technique is **6** The clusters seem to be congested with observations and are **overlapping** to each other except for cluster 1 and 6 which are distant with few numbers of observations. This may change due to number of clusters

PRINCIPAL COMPONENT ANALYSIS (PCA)

A key aspect of the efficient performance of huge, high-dimensional datasets is dimension reduction. It might be the main goal of data mining to analyze and visualize high-dimensional data, or it might be a necessary intermediary step to other analyses like clustering. A data reduction technique called principal component analysis transforms a large number of linked variables into a smaller set of linked variables called principal components.

The primary purpose of a principal component analysis is to reduce the number of dimensions in various artificial intelligence applications, such as computer vision and picture compression.




```
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6
Standard deviation     2.2628307 1.6282440 1.4140654 1.2399925 1.2262780 0.98069826
Proportion of Variance 0.3413602 0.1767452 0.1333054 0.1025054 0.1002505 0.06411794
Cumulative Proportion  0.3413602 0.5181054 0.6514108 0.7539162 0.8541668 0.91828470
                           Comp.7     Comp.8    Comp.9     Comp.10     Comp.11
Standard deviation     0.70997586 0.61993088 0.4056311 0.328832395 0.254326935
Proportion of Variance 0.03360438 0.02562095 0.0109691 0.007208716 0.004312146
Cumulative Proportion  0.95188908 0.97751003 0.9884791 0.995687854 1.000000000
```

Based on the **R** output, the maximum number of components is suggested to be **4**. For visualization, normally 3 components are advised but, in this case, more components are above the **90%** cut off point of the Principal Component Analysis. The sum of eigen values is equal to the total number of quantitative variables.

MULTIDIMENTIONAL SCALING:

A multivariate data analysis technique called multidimensional scaling (MDS) depicts sample similarity and dissimilarity by placing points on two-dimensional graphs. Multidimensional scaling in statistics is a technique for displaying the similarity of observations in a dataset in an amorphous cartesian space.

```
initial  value 23.983374
final  value 23.983355
converged
initial  value 11.250528
final  value 11.250504
converged
initial  value 1.109450
final  value 1.109439
converged
initial  value 0.815118
final  value 0.815108
converged
$points
               [,1]         [,2]         [,3]         [,4]
  [1,]  290.6859263  -47.6683552  -24.2777478   -5.96593044
  [2,] -248.7321745   34.0778878   94.6552693   27.38228230
  [3,] -303.1314528 -141.7626571   65.4161861  -28.19749548
  [4,] -261.1317682  134.9731678   67.6559651  -13.15853931
```

Based on the nature of the dataset, its better to go with the approach of **3** dimensions for visualizations the **4th** iteration goes to the cutoff point of **80%** the multidimensional scale. From the output, observation go through several iterations where each reaches a final value of convergence.

OUTLIERS:

Outliers make a data more variable, which reduces statistical power. Therefore, eliminating outliers can make the findings statistically significant. Some outliers in a dataset represent normal population variance and ought to be left alone. They are referred to as real outliers. Because they represent measurement errors, data input or processing flaws, or inadequate sampling, other outliers are harmful and ought to be eliminated.

From the **Supply Chain** dataset, a run through of mahanalobis distance method was performed and outliers were not spotted due to how congested and packed the observations are. This is a case of data analysis.

CANONICAL CORRELATION:

The relationships between two sets of variables are found and measured using canonical correlation analysis. When there are several intercorrelated outcome variables, however, and multiple regression would not be appropriate, one should instead use canonical correlation.

```
eigen() decomposition
$values
[1] 4.493775e-03 1.494390e-06

$vectors
           [,1]        [,2]
[1,]  0.2408958 -0.99785878
[2,] -0.9705510  0.06540528

eigen() decomposition
$values
[1] 4.493775e-03 1.494390e-06

$vectors
          [,1]       [,2]
[1,] 0.9576226 -0.4377802
[2,] 0.2880261  0.8990820
```

From the above output, the dataset was divided It is known that in the canonical correlation analysis is the square root of the maximum value of lambda between m1 and m2.

MEASURES OF ASSOSIATION:

CRAMER'S V

For the Supply chain dataset, Cramer's v was used depending on the variables, showing a strong relationship of 0.6. A chi-square test of independence effect magnitude is measured using Cramér's V. It assesses how closely two category fields are related. Using the following formula, the effect size is determined: Choose the field with the fewest number of categories. A chi-square test of independence effect magnitude is measured using Cramér's V. It assesses how closely two category fields are related.

REFERENCE.

Rich Huebner, PhD. Human Resources Data Set

https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis