Northwind Data Mining and Statistical Analysis Project

Steven Duchene

MIS 450 – Orenthio Goodwin

Colorado State University – Global Campus

June 7, 2020

Northwind Data Mining and Statistical Analysis Project

Academic essays should begin with an introduction. The introduction will provide readers with the context necessary for understanding your argument and the body of your paper. When composing the introduction, think about what context or background information the reader would benefit from knowing. Once your context is established, transition from that context into your thesis statement. The thesis statement generally comes at the end of your introduction and usually consists of a few sentences that sum up the argument for your paper overall. Thesis statements should also provide a roadmap for the reader so that they can navigate through the ideas present in the rest of your paper.

Analysis of the Variables

The Northwind Dataset is comprised of 22 different variables. Below, an analysis will be conducted on the individual variables. The variables will be presented alphabetically and, where appropriate, the summary statistics will be provided and discussed.

1. Category ID: The category ID is a numerical value which is assigned to the category dependent on the type of product. The value ranges from 1 to 8. This variable may serve as a good candidate for classification.
2. City, Country, and Region: The city, country, and region variables are character variables. They describe the location details as they relate to the purchasing customer. These variables will likely serve as good candidates for clustering. Additionally, in depth analysis can be conducted to determine the locations which perform best.
3. Customer ID: The customer ID is a numerical variable which is assigned to each customer. It is specific to the individual customer.

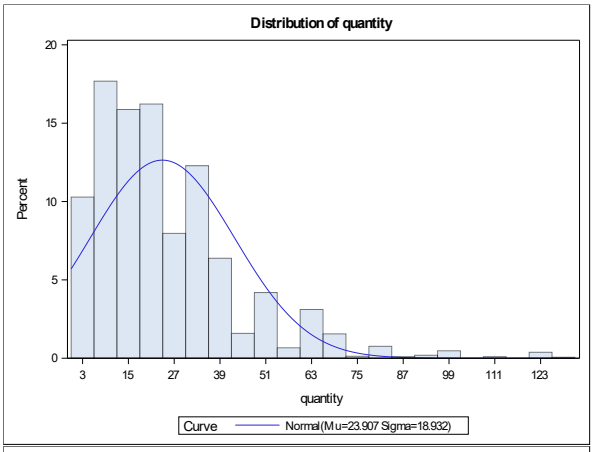
4. Discount: The discount variable is a numerical based variable. It is the determining factor for when a discount should be applied to a transaction. It has a minimum value of 0, a maximum of .25, a mean value of .056 with a standard deviation of .082.
5. Employee ID: The employee ID is a numerical variable which is assigned to each employee. It is specific to the individual employee and ranges from 1 to 9. The employee ID variable will serve to assist in the statistical analysis of employee performance.
6. Employee First and Last Name: These name variables are assigned to each employee ID and serve as an identifying value for the employee.
7. Freight: The freight variable is a numerical value. This value is used to determine the freight cost associated with each order and its accompanying shipping details.
8. Order Date: The order date variable is of the data type value. This variable specifies the date an order was made and is associated with the order ID. This variable could serve as a good candidate for clustering based on seasons.
9. Order ID: The order ID variable is a value assigned to each order and is specific to each order in the database.
10. Product ID: The product ID variable is a value assigned to each product in the database. It is specific to each individual product and is associated with the product name. The product ID could serve as a good candidate in cluster analysis. There are 76 different products.
11. Product Name: The product name is the name assigned to the individual products in the database. It is associated with the product ID.
12. Quantity: The quantity variable is associated with the order ID and product ID. It specifies the number of product ordered with respect to each individual order.

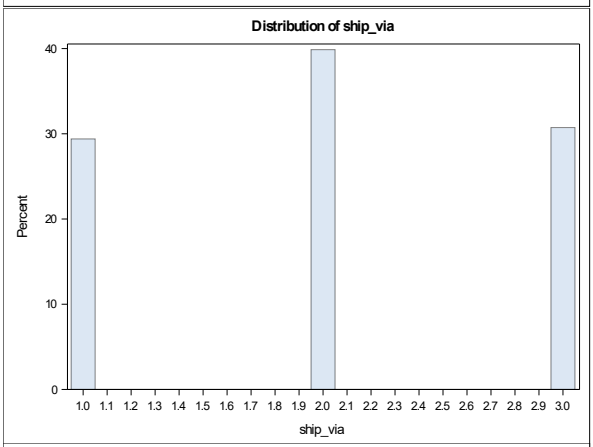
13. Quantity per Unit: The quantity per unit variable is specific to each product and its associated product ID. This variable specifies the quantity per unit ordered.
14. Required Date: This variable is of the date type and specifies the date the product is required by the customer. It is associated with the order ID and other variables.
15. Ship Via: The ship via variable specifies the method of shipping associated with the order ID. It is a numerical variable with three options. It may serve as a good candidate for classification.
16. Shipped Date: This variable is of the date type and specifies the date that an order was shipped. It is associated with the order ID.
17. Unit Price: The unit price variable is a numerical variable which specifies the cost per unit of each product. This variable is associated with the product ID. It has a range of values equaling 261.5 with a mean value of 25.60 and a standard deviation of 29.28.

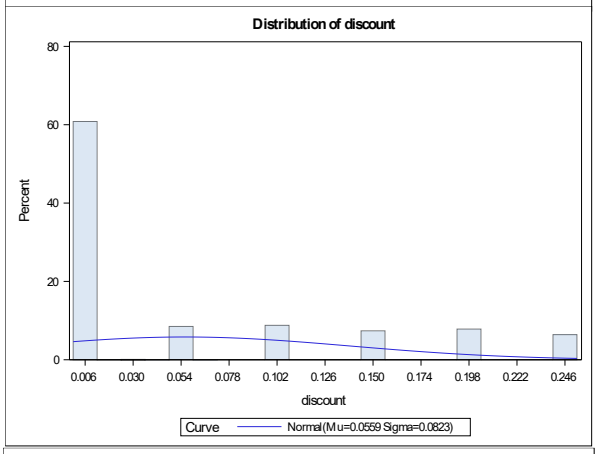
Statistical Analysis of the Appropriate Variables

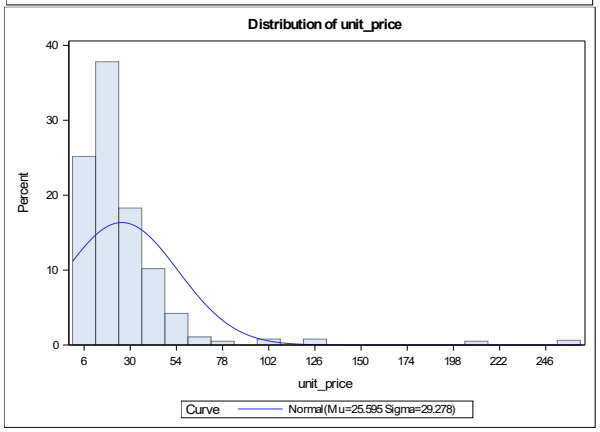
Due to the variability of the data in the Northwind Dataset, not all variables require summary statistical analyses, nor would they likely make sense. Where appropriate, summary statistics were calculated for specific variables. Where appropriate, alternative analyses will be conducted. They are presented below.

Variable	Mean	Std Dev	Minimum	Maximum	N	Range
ship_via	2.0133333	0.7754244	1.0000000	3.0000000	3150	2.0000000
quantity	23.9073016	18.9324897	1.0000000	130.0000000	3150	129.0000000
discount	0.0559333	0.0823238	0	0.2500000	3150	0.2500000
unit_price	25.5954508	29.2778283	2.0000000	263.5000000	3150	261.5000000
freight	94.0552730	132.3345100	0.0200000	1007.64	3150	1007.62

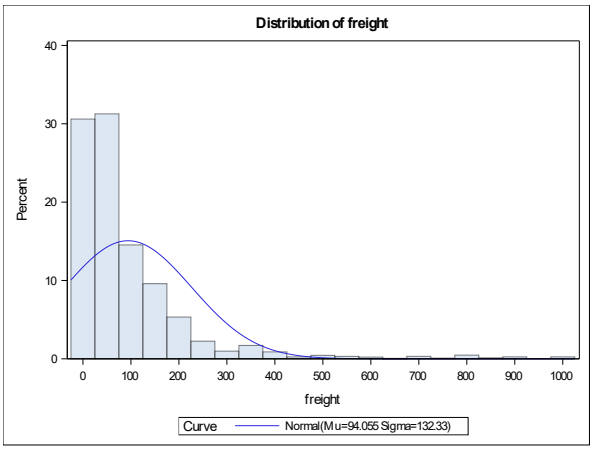






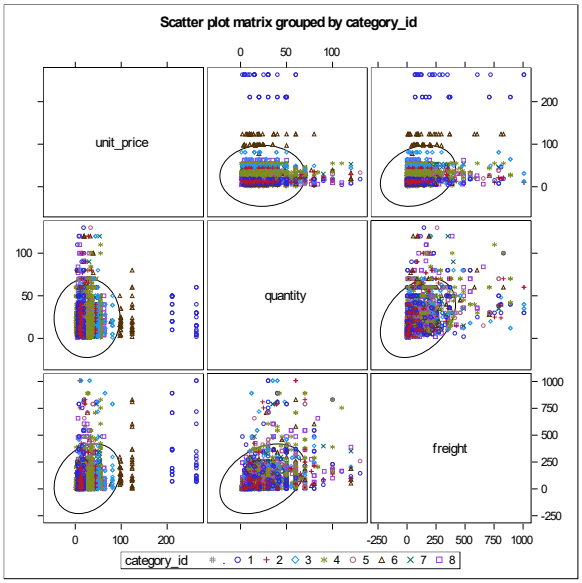


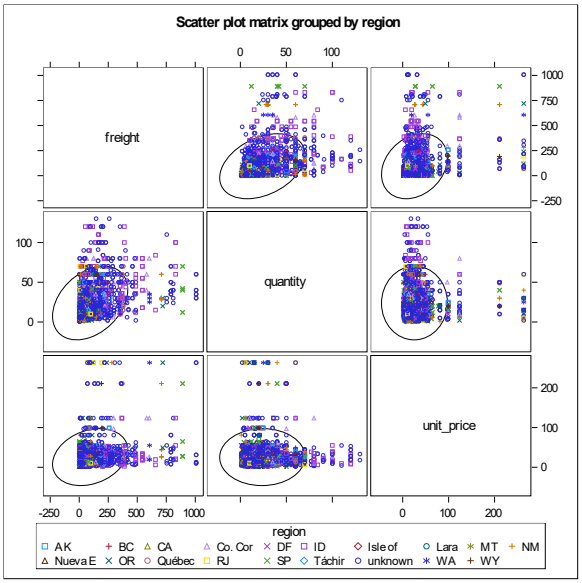
- Quantity: Based on the summary statistics above, the average quantity of items ordered was 23.91 with a range of 129, and a standard deviation of 18.93. The data exhibits a right skew as indicated below.
- Ship Via: As is observed in the histogram to the left, options 1 and 2 were utilized approximately 30% of the time; respectively. Option 3, the most popular option, was utilized nearly 40% of transactions,
- Discount: Based on the summary statistics above, the average discount was .056 with a range of .25, and a standard deviation of .082. It is not normally distributed and appears to have a strong right skew.
- Unit Price: Unit price has a mean value of 25.6, a range of 261.5, and a standard deviation of 29.28. It is strongly skewed to the right as indicated by the histogram and normal distribution overlay.
- Freight: Freight has a mean value of 94.06, range of 1007.62, and standard deviation of 132.33. It is skewed to the right.



Selection of Appropriate Classification Variables

Statistical analysis is an important part of the data mining process. In order to effectively determine the appropriate analyses to be completed, classification tasks must take place. Classification is perhaps the most common data mining task encountered in the field (Larose, 2015). It includes techniques such as k-Nearest Neighbor, Decision Trees, neural networks, and logistic regression. One method of determining which variables could serve as appropriate classifiers is to use the Data Exploration tool in the SAS platform. After conducting several iterations of data exploration, several variables were identified as good candidates for classification variables.





Based on the scatter plot matrix to the right, the region variable appears to be a good candidate for use as a classifier. When compared to the quantity and freight variables, a slight linear relationship was observed. A similar relationship was observed with unit price. Another

variable that was discovered as a possible candidate for use as a classifier was the category ID variable. Using the same variables as above, a similar, linear relationship was observed. Additional data exploration tasks were completed but did not yield similar results.

Cluster Analysis

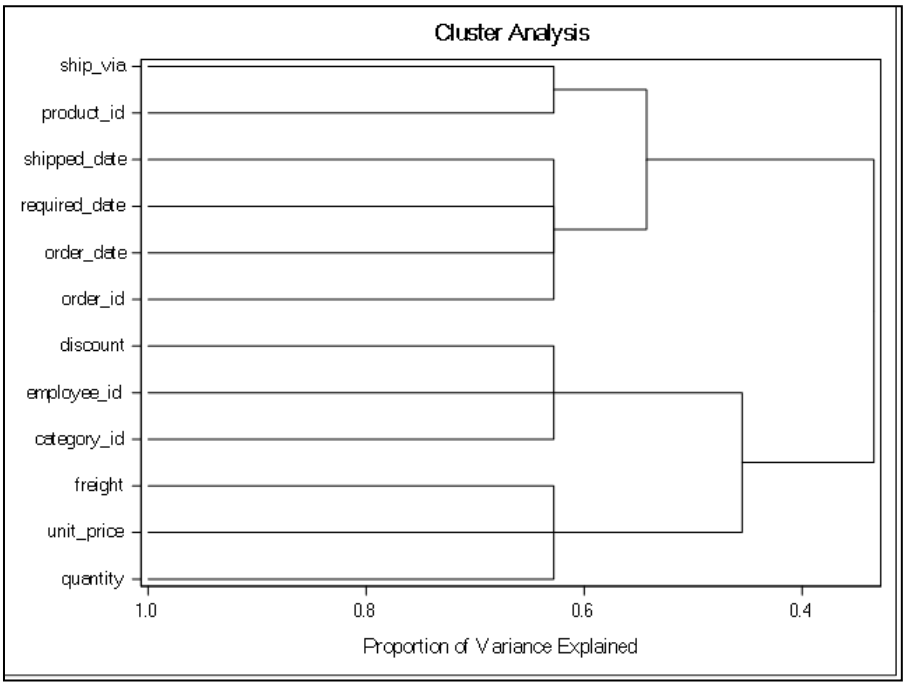
Unlike other data comparison methods, such as the classification task discussed above, cluster analysis attempts to group data based on the similarity of whole cases rather than individual variables (Leonard et al, 2008). There is a critical need for organizations to gain a deeper understanding of the nature and characteristics of their customer base. There are several methods that can be completed in order to conduct cluster analyses on the dataset in question. These methods include hierarchical, k-means clustering, and birch clustering (Larose, 2015). In order to determine whether or not the variables in the dataset would be appropriate for clustering, several tasks were completed using the SAS statistical platform. These methods included the "cluster variables" and "k-means clustering" options in SAS.

For the k-means clustering observation, several iterations were completed in order to select the ideal number of maximum clusters. In order to achieve this, the Pseudo-F Statistic was utilized. The F-Statistic has been identified as one of the top criteria to determine the appropriate number of clusters (0, 2019). At around 10 clusters, the F-statistic started to deteriorate. In all honesty, it continued to increase as the number of clusters decreased. Based on an analysis of theses statistics, a value of 6 clusters was selected. The initial seeds chart is below.

Initial Seeds							
Cluster	**product_id**	**category_id**	**employee_id**	**order_id**	**order_date**	**required_date**	**shipped_date**
1	0.881578947	0.285714286	0.000000000	0.984318456	0.992548435	0.981077147	0.996992481
2	0.973684211	0.000000000	0.125000000	0.038600724	0.061102832	0.071324600	0.096240602
3	0.460526316	1.000000000	0.875000000	0.436670688	0.575260805	0.573508006	0.589473684
4	0.026315789	0.142857143	0.250000000	0.352231604	0.475409836	0.475982533	0.508270677
5	0.460526316	1.000000000	1.000000000	0.008443908	0.011922504	0.023289665	0.007518797
6	0.486842105	0.000000000	1.000000000	0.773220748	0.882265276	0.873362445	0.891729323

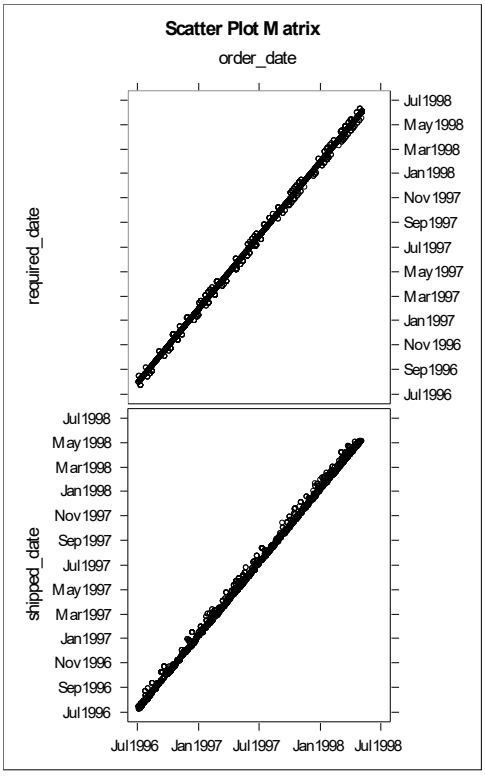
Initial Seeds					
Cluster	**ship_via**	**quantity**	**discount**	**unit_price**	**freight**
1	0.000000000	0.418604651	0.000000000	0.040152964	0.029842599
2	0.000000000	0.224806202	0.000000000	0.016061185	0.008892241
3	0.000000000	0.155038760	1.000000000	0.065009560	0.026557631
4	1.000000000	0.457364341	0.000000000	0.030592734	1.000000000
5	1.000000000	0.186046512	0.000000000	0.050478011	0.147188424
6	1.000000000	0.302325581	0.000000000	1.000000000	0.278468073

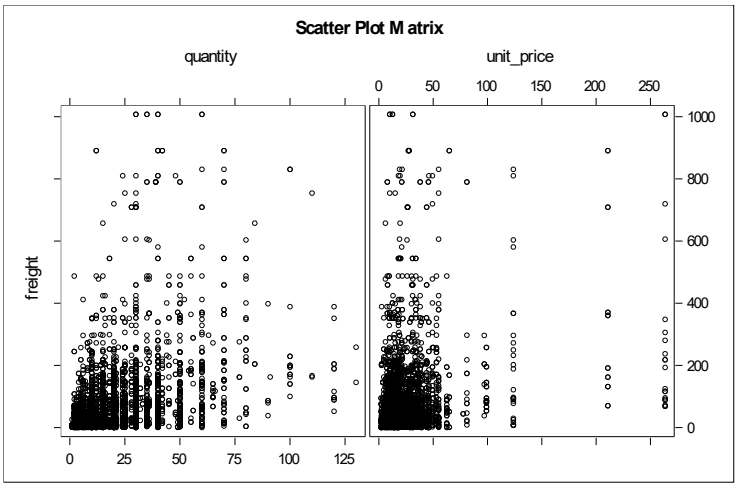
When compared to a general cluster of the variables, completed in SAS, some patterns and some differences are observed. The cluster obtained from SAS is presented below. The basic cluster analysis was able to cluster variables which made sense; for the most part. For example, cluster four is comprised of "freight, unit price, and quantity." These are likely clustered together due to their characteristics. As unit price and quantity increase, it is likely that the cost to ship the items also increases. What I found interesting, however, was that k-means clustering produced very different results. This is likely attributed to a level of incompatibility of the dataset for that type of clustering model.



Associations Between Variables

In order to observe possible associations between variables, a basic correlation analysis was completed using the SAS statistical platform. Correlation analysis deals with the relationships between variables and is a measure of linear association between two variables (Tutorials Point, n.d.). The analysis returns values between -1 and 1. As values near the positive and negative extremes, the analyst is able to determine whether or not a relationship exists; be it positive, negative, or not at all. As expected, variables such as order date, shipping date, and required date exhibited strong correlation. A weaker yet similar pattern was seen between quantity, unit price, and freight. Outside of the variables presented here, little to no correlation was observed.





Recommendations and Conclusions

After an extensive analysis and review of many different characteristics of the dataset, there were several conclusions that could be made. Firstly, the region variable appears to be a good candidate for use as a classifier. When compared to the quantity and freight variables, a slight linear relationship was observed. A similar relationship was observed with unit price.

Another variable that was discovered as a possible candidate for use as a classifier was the category ID variable. It would be interesting to utilize these potential classifier variables to conduct further statistical analyses and data collection. Secondly, cluster analysis yielded results that made sense. However, it also became apparent that the dataset was not the best candidate for conducting k-means clustering. Finally, the dataset does appear to be suitable for meeting the organizations business goals. However, when looking back on the creation of the data warehouse and its associated fact and dimension tables, more analysis should be conducted to ensure that the appropriate and most usable data is obtained from the database. In doing so, it will ensure that the organization can capitalize on data, its associated trends, and its predictive abilities.

References

Leonard, S. T., & Droege, M. (2008). The uses and benefits of cluster analysis in pharmacy research. *Research in social & administrative pharmacy : RSAP, 4*(1), 1–11. https://doi.org/10.1016/j.sapharm.2007.02.001.

Larose, D. T., & Larose, C. D. (2015). *Data mining and predictive analytics*. Hoboken (NJ): Wiley.

Tutorials Point. (n.d.). SAS - Correlation Analysis. Retrieved from https://www.tutorialspoint.com/sas/sas_correlation_analysis.htm.

0, M. (2019, January 28). 10 Tips for Choosing the Optimal Number of Clusters. Retrieved from https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92.