Project Milestone-Module One: Option One

Richard Klein

MIS581 – Enterprise Performance Management

Colorado State University – Global Campus

Dr. Mamdouh Babi

February 13, 2022

**Portfolio Project-Module Eight: Option Two**

Portfolio Project-Module Eight: Option Two, entitled, “Business Analytics in SAS Studio (Instructure, 2022)” reviews the analysis of the Orders\_Northwind dataset to determine how the organization might increase the quantity and profitability of their orders. The analysis will review whether there are strong predictor variables that are related to sales amount that can be used for these predictions (Instructure, 2022). Deliverables for this Portfolio Project include a description of the organizations business problem and strategic goals, the business questions asked and how they will solve this problem and achieve the organization’s strategic goal, a null and alternate hypothesis for each of the business questions, the descriptive and statistical tests used to include charts, graphs and tables, analysis of findings, and recommendations for further analysis (Instructure, 2022).

**Business Problem and Strategic Goals**

Northwind Traders is a fictitious company which imports and exports specialty foods from around the world (Yugabyte, 2022). The sample data was created by Microsoft and used for database tutorials for decades (Yugabyte, 2022). The dataset includes the variables order id, customer id, employee id, territory id, product id, category id, unit price, quantity, discount, gross sales, discount amt, and net sale.

The business problem for Northwind Traders is that they would like to better predict future business growth (Instructure, 2022). The strategy goals of Northwind Traders are to increase the quantity and profitability of their orders (Instructure, 2022). Finding relationships within the Northwind Traders dataset may provide the answers to business questions posed below.

**Business Questions and Justifications**

The following business questions were derived by reviewing the Northwind Trader’s dataset. The dataset provides sales and discount data for products provided by the company.

**Table 1**

*Business Questions. Null and Alternate Hypothesis*

|  |
| --- |
| Business Question One – Bigger discounts can lead to bigger or more frequent sales. Did higher discounts produce higher sales amongst customers?  H0 - Null Hypothesis – The customers with the highest percentage of discounts accounted for the top five customers in sales.  Ha - Alternate Hypothesis – The customers with the highest percentage discounts did not account for the top five customers in sales. |
| Why? The answer to this question could give the organization insight into whether they need to overhaul their discount program to raise sales amongst customers with lower sales figures. |
| Business Question Two – Employees may be underperforming in sales of certain products. Are there any statistically significant (10%) differences in percentage sales by product for each employee?  H0 - Null Hypothesis –There are statistically significant differences in percentage sales by product for each employee.  Ha - Alternate Hypothesis –There are no statistically significant differences in percentage sales by product for each employee. |
| Why? Incentivizing employees to improve sales in product areas that are underperforming can help to increase sales. |
| Business Question Three – Products may be underperforming in sales regions. Are there any statistically significant (10%) differences in product sales by region?  H0 - Null Hypothesis –There are statistically significant differences in product sales by region.  Ha -Alternate Hypothesis–There are no statistically significant differences in prod sales by region. |
| Why? Incentivizing employees to improve sales of products in underperforming regions can increase sales. Incentivizing customers by offering discounts on products in regions that are underperforming can increase sales. |
| Business Question Four – Customers may not be offering all available product categories to customers. Are there any product categories that customers are not offering (0% sales) their customers that we could incentivize them to purchase?  H0 - Null Hypothesis –There are product categories that are not being offered by certain customers to their clientele.  Ha - Alternate Hypothesis – There are not product categories that are not being offered by certain customers to their clientele. |
| Why? Incentivizing customers to add additional product categories to their inventory could boost the sales of several products. |

*Note*. Four business questions related to Northwind’s dataset adapted from "Hypothesis Testing" by StatisticsHowTo.com, 2022. Copyright 2022 by Statistics How To.

**Descriptive Statistics**

To begin, the SAS test Proc Means was used to generate descriptive statistics for six quantitative variables in the Northwind’s dataset. There variables tested under Proc Means are unit\_price, quantity, discount, gross\_sale, discount\_amt, and net\_sale. Some numbers to take note of are- unit prices which can range from between $2 and $99 with an average of $20. Quantity of goods ordered can range from 1 to 70, with 24 being the average. Discount average is .0592520 percent with a standard deviation of .0916635. The Gross sale average is $492.97 while the Net sale average is $456, with the average discount amount of $36.

**Figure 1**

*Proc Means descriptive statistical test*

Graphical user interface, application

Description automatically generated

To get additional descriptive statistics which examine the distribution of data from the Northwind dataset the Proc Univariate test was used on the same variables from above unit\_price, quantity, discount, gross\_sale, discount\_amt, and net\_sale (Guido, 2009). For example Figure 2 below shows the distribution of the Unit Price which previously mentioned has an average (mean) unit price of $20, but the mode shows that the unit price most frequently tallied is $15.20, and the distribution in the figure shows that the highest counts are between six dollars and $18, with a skewness of 2.5708 which means a tail to the right (towards higher prices), and kurtosis of 9.51 which means a distribution which is more peaked with heavier tails, such as shown in the Figure 2 image below.

**Figure 2**

*Proc Univariate Distribution for Unit Price*

Graphical user interface

Description automatically generated

Other descriptive statistics of note from the Northwind dataset using the Proc Univariate descriptive statistical test includes a discount Mode of zero, which means that is the most frequently applied discount, which applies to discount amount as well. Quantity has a normal distribution with positive skewness, with most frequent quantity (mode) being 15 units. Both gross and net sales also have a positive skewness, so a tail to the right towards higher sales numbers.

In addition to Proc Means and Proc Univariate, Proc Freq was used to create two- way tables to analyze multiple pairs of variables and discover relationships prior to using predictive analytics such as Proc Corr and Proc Reg. As there are a myriad of relationships that can be analyzed, this analysis was geared towards the business questions asked previously. So, tables were built to study the relationship of discounts and customer id, employee id and product id, and territory id and product id.

**Figure 3**

*Two-way table for customer id and discount percentage*

*A screenshot of a computer

Description automatically generated*

The first relationship that was examined (discounts and customer id) shows that the customer that had the most orders also had the most discounts applied to their sales (34 orders total with 25% discount on 21 orders and 20% discount on nine orders). This did not hold true for the second highest order holder who had no discounts applied. However there does appear to be a correlation as the third and four highest order holders do have discounts applied to a considerable number of orders. This led to an examination of customer id and products, and products and discounts. Of note are that product id’s 2, 16, 24, 74 are the highest selling products and all but product id two had discounts applied in at least half of its sales.

The second relationship examined with a two-way table was employees and the products that they sold. This two-way table examined in Figure 4 below shows that all but employee nine and in some cases employee five underperformed on selling (greater than 10% difference) the most sold products 2, 16, 24, 74, and 41. An examination of discounts applied by employee shows that the top four employees by number of sales applied discounts half of the time.

**Figure 4**

*Employee id and product id examined*

A screenshot of a computer

Description automatically generated with medium confidence

The third business question required the examination of territory id and product id in frequency tables to determine if there existed differences in sales of products amongst different areas. The table clearly shows that the highest selling products for Northwind were not sold in a considerable number of territories so there exists opportunity to expand product offerings in different territories.

**Figure 5**

*Territory and the products sold*

*A screenshot of a computer

Description automatically generated with medium confidence*

The final business question related to territories and the product categories that are sold. The examination of this data yielded even more statistical evidence that there exist opportunities to expand category offerings to different territories as shown in Figure 6 below where every territory has a category of product that is not beings sold.

**Figure 6**

*Territory and Category of products sold*

*A screenshot of a computer

Description automatically generated*

**Predictive Statistics**

This section focuses on Predictive analytics tests on the Northwind dataset in SAS. The procedures Proc Corr and Proc reg will be studied to examine the strength of the relationships between the variables in the dataset to confirm previous analytics and concentrate organizational focus for answering their business problem of increasing quantity and profitability of their sales (Datafair, 2022). Figure 7 below shows the initial Proc Corr procedure run for all variables within the Northwind dataset. Pearson Correlation shows significant correlation between a few variable pairs, such as discount amount and gross sales with a .68 correlation, gross sales and unit price with a .62 correlation, discount amount and quantity with a .66 correlation, net sale and unit price with a .64 correlation.

**Figure 7**

*Correlation Table*

Graphical user interface, application, table, Excel

Description automatically generated

Further Proc Corr procedures and linear regression using Proc Reg procedures were run between variables that had higher correlations. Data generated from procedure such as the scatter plot below in Figure 8 and the Fit Plot for linear regression in Figure 9 show the positive relationship between discount amount and gross sales.

**Figure 8**

*Scatter plot showing relationship between discount amount and gross sales*

Graphical user interface, chart

Description automatically generated

**Figure 9**

*Linear Regression Gross Sales and Discount Amount*

Graphical user interface, chart, scatter chart

Description automatically generated

Similar procedures were run for discount amount and quantity which showed a high correlation, and discount and quantity which did not show a high correlation, possibly meaning that higher quantities of products purchased showed discount amounts, but there wasn’t necessarily a strategy or a fixed discount program where a higher quantity of products purchased led to a higher discount percentage which may be an opportunity for the organization.

Finally, automated model selection was completed using Proc Reg and the backward elimination technique. The dependent variable was gross sales and the independent variables remaining after elimination which were significant at the .1000 level were unit price, quantity, discount, and discount amount.

**Figure 10**

*Automated Model Selection*

Graphical user interface

Description automatically generated

**Analysis**

After completion of procedures to review descriptive and predictive analytics for Northwind’s dataset, there is a need to review the business questions posed and whether the analytics provided any answers. Business question one asked, “Bigger discounts can lead to bigger or more frequent sales. Did higher discounts produce higher sales amongst customers? The null Hypothesis which states that the customers with the highest percentage of discounts accounted for top customers in sales was correct based on the data from the frequency table. Additionally, one of the highest Pearson correlation scores (.68) existed between the variable sales and discount amt and linear regression showed a strong relationship to make this paring a focus of the organization’s strategy moving forward.

Business question two asked, “Are there any statistically significant (10%) differences in percentage sales by product for each employee?”. The null Hypothesis which states that there are significant differences in percentage sales by product for each employee was correct. Evidence for this was in the two-way frequency table for employee id and product id, which showed that 3 of 9 employees sold 100% of the most sold product, while one of the nine employees sold 100% of the second most popular product. One employee (employee ID 9) accounted for over 50% of sales of top four products, while four employees sold none of those products.

Business question three asked, “Are there any statistically significant (10%) differences in product sales by region?”. The null Hypothesis which states that there are statistically significant differences in product sales by region was correct. Again, a two-way table showed that sales of products by region were significantly different. 18 of the territories did not have any purchases of the top four products sold by Northwind’s in this dataset’s period.

Business question four asked, “Are there any product categories that our customers are not offering (0% sales) their customers that we could incentivize them to purchase?”. The null Hypothesis which states that there are product categories that are not being offered by our customers to their clientele is correct. Analysis shows that based on product offerings in all the different territories there is at least one product category per territory that is not being offered. An analysis by customer id and category shows that each customer has at least two category of products that they did not buy and offer to their clients.

**Recommendations**

As the data and analysis indicates that there is a clear correlation between discount amount and gross sales, but not a clear correlation between quantity and discount percentage, the organization should look to implement a transparent product discount program based on product quantity purchased to generate additional sales. This as well as incentivizing their sales staff to sell their most popular products in all their territories should increase quantity and profitability of sales as well as better predict future business growth. Further analysis should review what products are being offered, and why not all customers carry said items, sales by customer, region, and territory.

**Conclusion**

Portfolio Project-Module Eight: Option Two, entitled, “Business Analytics in SAS Studio (Instructure, 2022)” reviewed the analysis of the Orders\_Northwind dataset to determine how the organization might increase the quantity and profitability of their orders and better predict future growth. Using descriptive and predictive analytics to determine predictor variables related to the organizations sales it was determined that there was a high correlation between discount amt and sales. There was also no clear discount percentage strategy that could be found. Therefore, a recommendation was given to enact a discount program based on product units sold, as well as expand offerings to all customers in all regions through their sales reps so that the organization can increase the quantity and profitability of sales and better determine future growth.

Just a final note regarding the portfolio project. I found it to be challenging but an extremely rewarding exercise in the use of SAS and procedures to generate descriptive and predictive analytics from a dataset. Thank you for your time.

**Reference**

Dataflair.com. (2022). SAS Correlation Analysis – Understand the PROC CORR & Correlation Matrix. Retrieved from https://data-flair.training/blogs/sas-correlation-analysis/.

Elliott, A., & Woodward, W. (2016). SAS Essentials: Mastering SAS for Data Analytics. (2nd ed.) John Wiley & Sons, Inc.

Guido, J. (2009). Guido’s Guide to PROC UNIVARIATE: A Tutorial for SAS® Users. Retrieved from https://www.lexjansen.com/nesug/nesug09/sa/SA07.pdf.

Instructure. (2022). Module Three: Portfolio Milestone. Retrieved from https://csuglobal.instructure.com/courses/55615/assignments/1102810?module\_item\_id=2563984.

Statistics How To. (2022). What is Chi-square Statistic? Retrieved from https://www.statisticshowto.com/probability-and-statistics/chi-square/

Statistics How To. (2022). Hypothesis Testing. Retrieved from https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/.

Yugabyte. (2022). Northwind Sample Database. Retrieved from https://docs.yugabyte.com/latest/sample-data/northwind/.