Portfolio Project Option #2

Daniel Luka

MIS543: Enterprise Performance Management

Colorado State University – Global Campus

Dr. Mamdouh Babi

June 14, 2020

Table of Contents

1. Project Background...…………………….…………………………….……….….………….3
   1. Business Problem…...……….………………………………………..……….….…...3
   2. Data Overview...…………….…………………………………..….……………...….3
   3. Business Questions and Hypotheses.………………………………………………….4
2. Descriptive Statistics…………………………………………………………………………..5
   1. Proc Corr………………………………………………………………………………5
   2. Proc Univariate………………………………………………………………………..6
   3. Proc Sgscatter…………………………………………………………………………8
   4. Proc Sgplot………………………………………………………………………….…9
3. Predictive Statistics…………………………………………………………………………..10
   1. Proc Reg..…………………………………………………………………………….11
   2. Proc Anova…………………………………………………………………………...12
   3. Proc Rsquare..………………………………………………………………………..13
4. Analysis………………………………………………………………………………………14
5. Recommendations……………………………………………………………………………15
6. Conclusion…………………………………………………………………………………...15

Retail Distribution Company Analysis

**1 Project Background**

A retail distribution company is seeking analytic assistance focused on its order data (CSU global., n.d. b). Within the analytics request there are two distinct intents, explanation and forecast. First, provide insights into quantity and profitability maximization (CSU Global, n.d. b). Generally, this component will necessitate the use of descriptive statistics to explore the underlying data. Second, strengthen the organizations predictive capacity, particularly as it relates to sales and quantity enlargement (CSU Global, n.d. b). The following analysis begins with business problem derivation, data overview, and business questions / hypothesis summary (CSU Global, n.d. b). Then, the problem, questions, and hypotheses will be analyzed with SAS descriptive and predictive analytics, to provide recommendations to company leadership (CSU Global, n.d. b). Note, the associated code and full result sets are provided alongside this word document as SAS and PDF files, respectively.

* 1. **Business Problem**

Clearly, leaders currently lack an understanding of the fundamentals of their business operations and financial health. Accordingly, they necessitate assistance in deriving a strategy to comprehend, grow, and forecast their business.

**1.2 Data Overview**

Leadership has provided the orders\_northwind\_mod8.csv dataset to form the foundation of the analytic task at hand. Prior to declaring business questions and running statistical tests on the data, an initial review of attributes and observations is advisable (Han, Kamber, & Pei, 2012). Said file contains 254 records that are comprised of 13 identification and financial attributes. The categorical identifiers include order\_id, category\_id, customer\_id, employee\_id, territory\_id, and product\_id. For the sake of standardization, a numeric version of the customer\_id, customer\_id\_num, was added with excel by assigning a unique number to each customer. The financial attributes include unit\_price, quantity, discount, gross\_sale, discount\_amt, and net\_sale. Net\_sale will represent profitability, as it is gross\_sale less applicable discount\_amt.

**1.3 Business Questions and Hypotheses**

The logical starting point, given the first aim of this project, is a comparison business question surrounding financials (CSU Global., n.d. a). Business question one (BQ1) is the following: Does profitability or volume differ across products for this distribution company? The null hypothesis would posit that no product stands above the rest in either financial measure, while the alternative hypothesis would identify winning and losing products. This descriptive starting point will shed light on what specifically is occurring with product profitability (Sharda, Delen, & Turban, 2018).

The second business question (BQ2) is geography and distributional focused, because of the inclusion of a territory id (CSU Global., n.d. a). Do sales territories influence profitability or volume for the distribution company? The geographical null hypothesis would be no discernable financial difference across territories, while the alternative hypothesis would feature performing and non-performing territories that should be focused on and expanded or contracted in the future.

Relational business questions attempt to extract correlation and causation from the provided data (CSU Global., n.d. a). Are discounts associated with a greater volume of sales? The null hypothesis would indicate discounts are not related to sales quantities or dollars, while the alternative hypothesis would contribute discounts to greater or lesser sales. Discount analysis is an important focus for the third business question (BQ3) as it impacts individual product profitability, though if more product is sold as a result then the outcome is advantageous. The fourth business question (BQ4) is another relational example that focuses on employees. Do employees impact profitability or volume? For the null hypothesis to hold, employee\_id would have no bearing on either financial measure, where the alternative hypothesis could help trim underperformers and promote those exceeding expectations. In addition to financial improvement, these insights would be informative to future recruitment efforts.

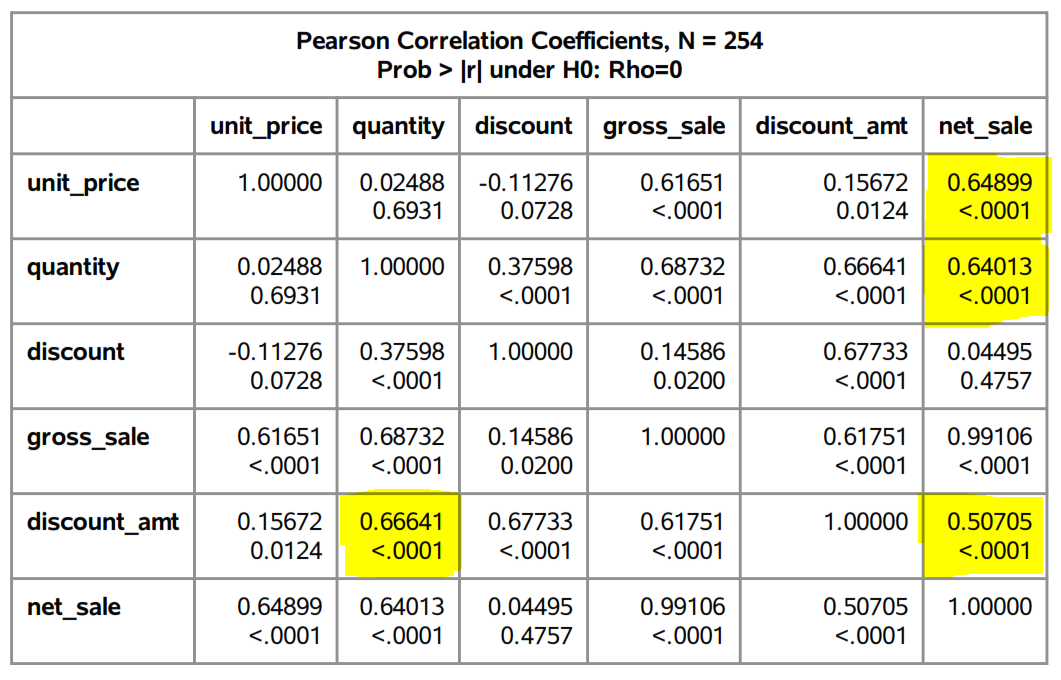
**2 Descriptive Statistics**

Descriptive statistics are the first statistical analysis post cursory data discovery. These tests precede predictive analytics to satisfy the need for more thorough understanding. Consequently, the following procedures were employed: proc corr, proc univariate, proc sgscatter, and proc sgplot. Justification for the procedures and a highlight of their results are provided in the individual sections below.

* 1. **Proc Corr**

Proc corr is the first statistical test performed post file ingestion, as it provides summary and relationship statistics for non-categorical financial attributes. Off the top, the data is representative of approximately $116,000 net sales and 6,000 products. The resulting average unit cost of $20 per product quantity is discounted 6% on average.

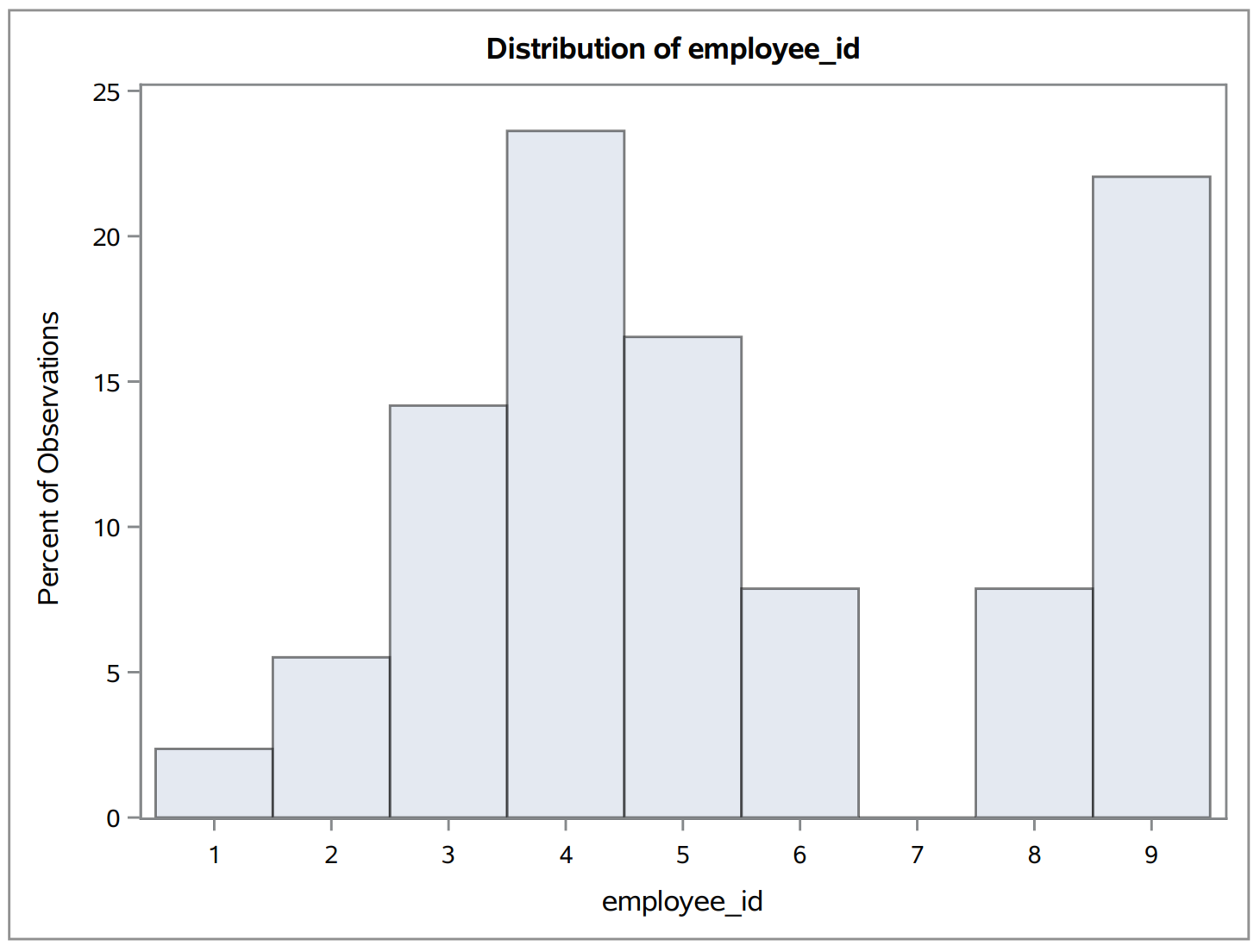
Following the summary statistics, the correlation matrix, in figure one below, shows the underlying relationships of the non-categorical variables. Here, net\_sale and quantity are the focus, given their near universal prevalence in the business questions. The strongest relationships for net\_sale, outside of the obvious gross sales attribute, are unit price, quantity, and discount\_amt. On the other hand, the standout for quantity is discount\_amt. The relationship between quantity and net sales, while apparent, is not the intent of any of the business questions and therefore will not be investigated further. Instead, the relationships between discount\_amt and both quantity and net\_sale, the focus of BQ3, are preliminarily suggesting an acceptance of the alternate hypothesis.



*Figure 1* Correlation Matrix. Adapted from CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

* 1. **Proc Univariate**

The next descriptive statistical method is proc univariate, which provides insights into the distribution of values within each of the thirteen attributes (Elliot & Woodward, 2016). This method offers the first glimpse at the categorical identifiers. Product\_id, the focus of BQ1, range from 2 to 77 and are relatively evenly distributed throughout the observations. Conversely, the attribute of choice for BQ2, territory\_id, can range from 1500 to 98000 and circa 85% of observations are territory ids from 1500 to 52000. The distribution of employee\_id from BQ4, in figure two below, range from one to nine. A lack of employee\_id seven in the observations means there are only eight employees represented in the data. Additionally, nearly half of the records are related to employees four and nine. This either indicates that the sampling method which produced this data set favored some employees over others or the relative sales performance of employees varies wildly. Unfortunately, without an understanding of how the dataset was derived, the uneven distribution of employees across observations will mean conclusions for BQ4 will require additional verification.



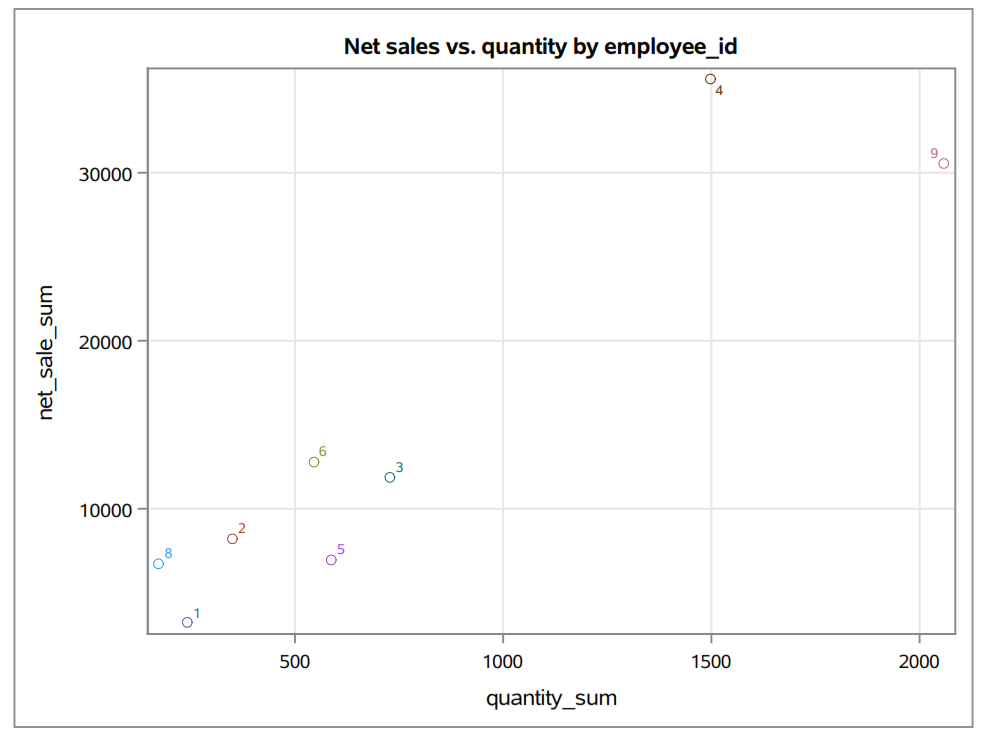
*Figure 2* Employee\_id Histogram. Adapted from CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

The financial insights from proc univariate are also informative. Particularly discount\_amt, the emphasis of BQ3, as nearly 75% of observations have less than a $25 discount applied. Another right-skewed distribution is net\_sale, as circa 50% of the observations are under $300, despite a maximum of nearly $3,000. Quantity, while still right skewed, is more normal than both discount\_amt and net\_sales. Ultimately, observations have a low discount\_amt and net\_sale in relation to their range, indicating both stand to be increased. While the distribution of attributes are interesting in isolation, the scatter and bubble plots that follow provide insights into the level of financial differentiation that exists with categorical variables.

* 1. **Proc Sgscatter**

The resulting scatter plots, from proc sgscatter, are concerned with the hypotheses for the categorical product\_id, territory\_id, and employee\_id associated with the first, second, and fourth business questions, respectively. Each scatter plot pits the net\_sale against quantity, with the datapoints representing the total for the categorical attributes of focus. To reduce the noise here and in the following bubble chart, proc\_sql was employed to produce data subsets that summarize the net\_sale and quantity for each categorical variable. The null hypothesis for each of the related business questions would hold if the datapoints were closely clustered in each of the scatter plots, as that would indicate similarity rather than difference.

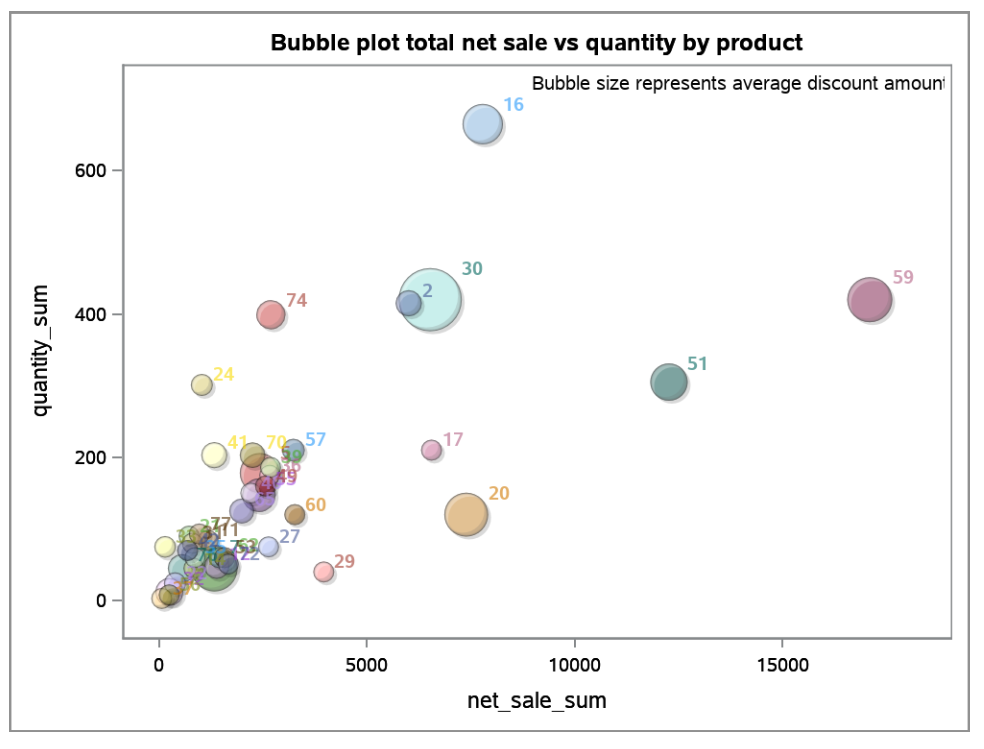
Fortunately, the opposite is true for all the categorical variables, preliminarily indicating acceptance of the alternate hypothesis for the first, second, and fourth business question. Product\_id is the most dispersed, though the standouts of 51, 74, 2, 30, and 16 are either above 400 total quantities sold or greater than $10,000 in total net sales. Territory\_id contains tighter groupings, though about four clusters could be identified. The top performing cluster contains three territories (27403, 20852, 27511) that all approach $12,000 in net sales and 500 product quantities. Finally, the employee\_id scatter in figure three below is the tightest clustering of the bunch. The three clusters here are employees with less than 500 quantities and $10,000 sold, less than 1,000 quantities and $20,000 sold, and greater than 1500 quantities and $30,000 sold.



*Figure 3* Employee\_id Scatter. Adapted from CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

* 1. **Proc Sgplot**

To round out the descriptive section, proc sgplot was used to produce bubble charts to satisfy the analysis of discount\_amt disparity for the third business question. Similar to the scatters, these graphs place categorical summaries on a graph of net\_sale vs quantity. What differentiates these from the scatters is the relative sizing provided by the average discount\_amt. Solidifying the acceptance of the alternate hypothesis for the third business question would rely on the bubbles sizing increasing alongside quantity and net\_sale on the graphs. The moderately strong relationship between discount\_amt and the categorical variables is displayed graphically in the bubble charts. In each case, there are outliers where large average discounts are associated with low quantities and net sales. Figure four shows the mixed results specific to product\_ids.



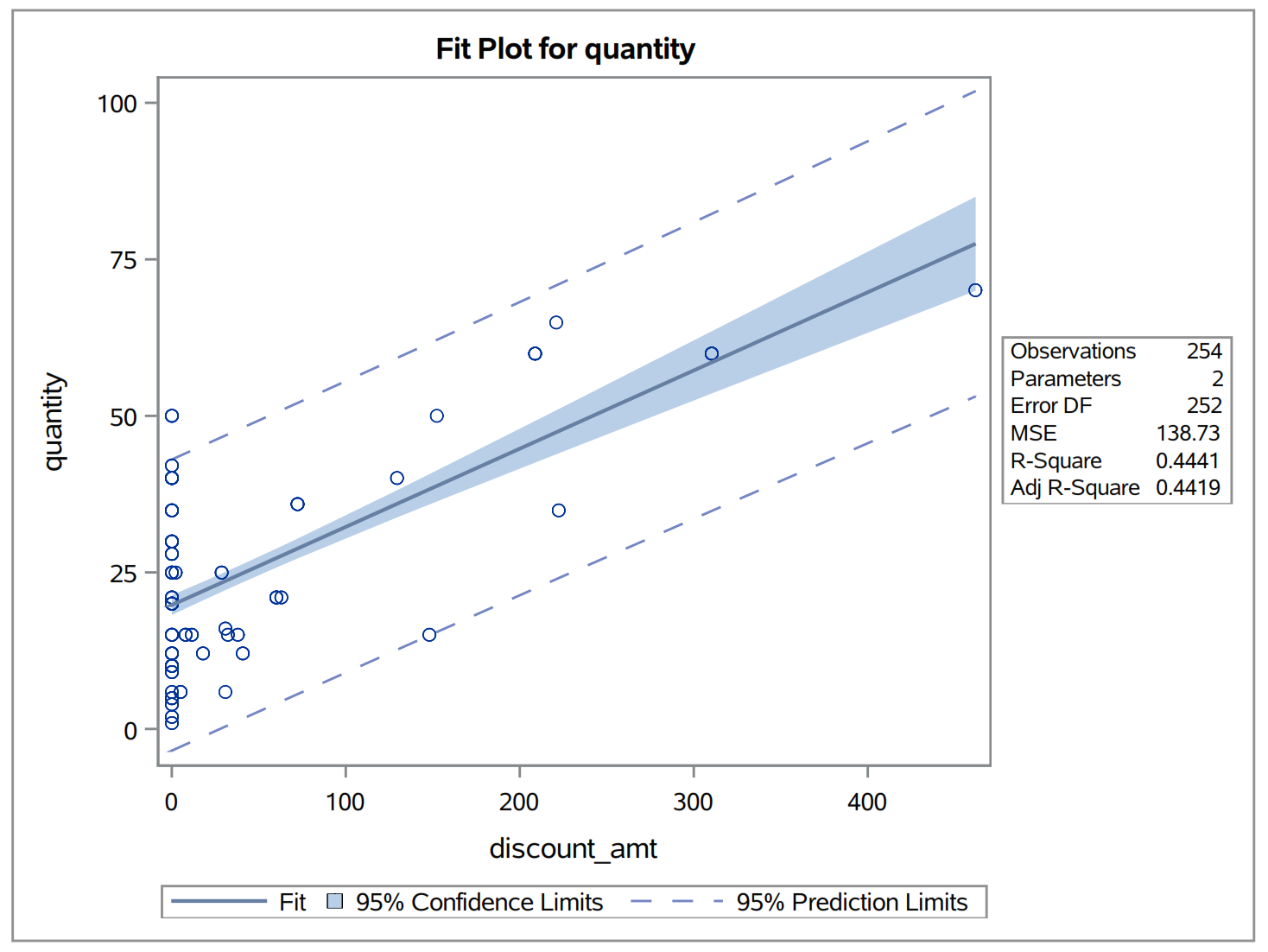
*Figure 4* Product\_id Bubble Chart. Adapted from CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

**3 Predictive Statistics**

Although the descriptive analysis has provided insights into the differences in categorical variables, prediction seeks to confirm the differentiators. Demand forecasting in retail distribution is reliant on descriptive classification to determine initial and replenishment shipments across the distribution network (Bradlow, Gangwar, Kopalle, & Voleti, 2017). As with description, differing predictive methods will be employed for the categorical and financial attributes. Proc ANOVA is utilized for categorical and proc reg is used for the financial independent variables (Elliot & Woodward, 2016). Lastly, proc rsquare is included to understand the value of possible attribute combinations.

* 1. **Proc Reg**

The scope of regression, limited to financial attributes, means the predictive insights are centered around discount\_amt for BQ3. The discount\_amt regressions for net\_sale and quantity produced r-squared values of .2571 and .4441, respectively. Essentially, this indicates discount\_amt can explain a quarter of the variance in net\_sales and almost half of the variance in quantity. A review of the fit plot in figure five for the strongest r-square of the two, between quantity and discount\_amt, shows a wide 95% prediction band that could benefit from more observations that contain a discount. Therefore, discount\_amt alone is not a good predictor of either financial metric, which would suggest the rejection of the alternate hypothesis for BQ3 to contrast the results of correlation.

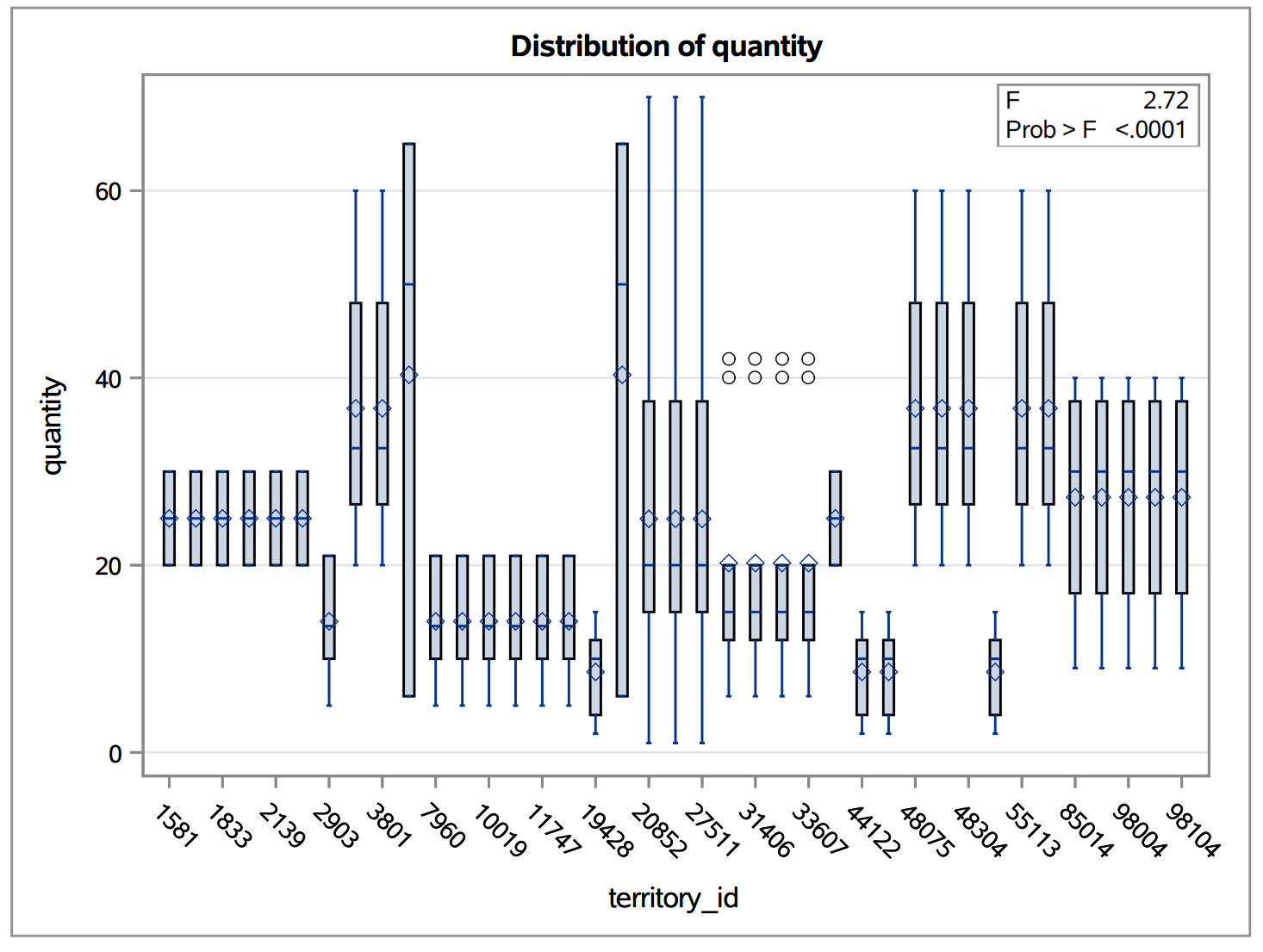


*Figure 5* Quantity vs Discount\_amt Fit Plot. Adapted from CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

Despite not specifically being associated with a business question, unit\_price was brought into regression analysis of net\_sale given the relationship identified within the correlation analysis. Combining the independent variables discount\_amt and unit\_price resulted in an r-square of 0.5896 and p<.0001, which is a significantly predictive relationship. Although outside of the scope of the business questions, the predictive capacity is noteworthy for the business problem at large.

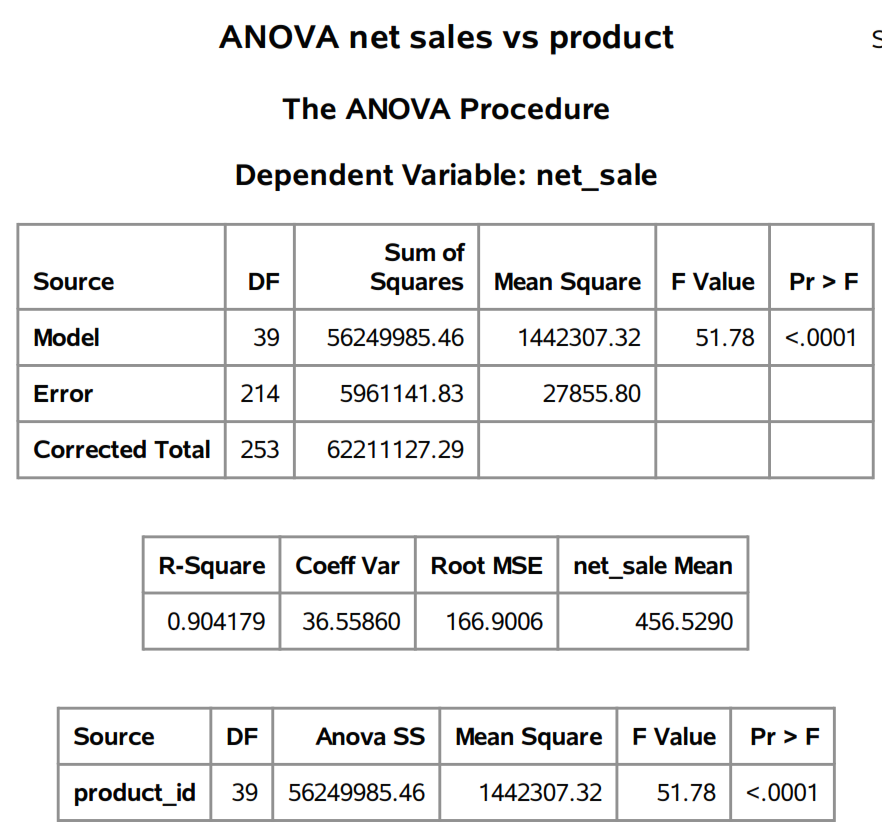
* 1. **Proc Anova**

Next, the predictive capacity of the independent categorical variables is tested with proc ANOVA. The output of proc ANOVA is comprised of tables of statistics and then box and whisker distribution graphs. The procedure was run twice for each independent variable, with net\_sale and quantity as the dependent variables, for a total of six tests. A particular standout of the plots, provided in figure six below, was the disparity among quantity across territory\_id. Further, the F value indicates the rejection of the null hypothesis for BQ2.



*Figure 6* Territory Box and Whisker Plot. Adapted from CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

To relate back to the regression analysis, the largest r-square statistical values for the ANOVA tests were .725 and .904 for product\_id vs quantity and net\_sale, respectively. Therefore, product\_id is a strong predictor for both financial metrics, accounting for more than 70% of the variation in both. The remaining statistics for the product\_id vs net\_sale ANOVA, including an exceedingly large F value of circa 51, are highlighted in figure seven.



*Figure 7* Product\_id ANOVA Statistics. Adapted from CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

* 1. **Proc Rsquare**

The last stop in the predictive analysis is proc rsquare, which combines variables to determine their consolidated variability capacity. Regrettably, this procedure did not work well with the included categorical variables, as the r-squares did not align with those in ANOVA. The only additional insight provided by this procedure in comparison to regression is that the discount attribute combined with discount\_amt and unit\_price improves the consolidated r-square value of discount\_amt and unit\_price alone, though not materially.

**4 Analysis**

Overall, the processes of data discovery, descriptive statistics, and predictive statistics was exceedingly informative. The resulting insights were plentiful enough to answer all the business questions. Perhaps the strongest results are for BQ1, where the alternate hypothesis was clearly accepted, rejecting the null hypothesis entirely. Both profitability and volume are substantially different across product\_id. Also, product\_id serves as the strongest categorical predictor of said financial metrics, with the highest r-square values by far.

Similarly, the alternate hypothesis is accepted for BQ2 and BQ4, despite both having limited predictive capacity. For BQ2, territory\_id was the second most dispersed across the scatters of net\_sale vs quantity. Additionally, the box and whisker plot of quantity by territory\_id is supportive. Related to employee\_id, high, medium, and low performing employees are visible on the scatter plot. A unique limitation of employee\_id is the uneven distribution of values, which could be due to sampling or different sales volume.

The contrarian results are associated with BQ3, as the null hypothesis is accepted. The problems here started with the value distributions. A majority of the observations have little to no discount\_amt, which makes it hard to prove the alternate hypothesis of an association with the financial metrics. The strongest regression relationship with discount\_amt still only accounted for less than half of the variance in quantity.

**5 Recommendations**

Importantly, this analysis is not meant to occur once in isolation, as indicated by the iterative recommendation of the Cross Industry Standard Process for Data Mining (Bradlow et al., 2017). Within the next iteration of the analysis, findings from this version can be strengthened or challenged alongside a review of new questions. Further, data readiness can be improved through the mitigation of concerns around data richness, meaning more attributes should be sourced, and consistency, which relates to the concerns of data sampling noted for the discount\_amt and employee\_id observation distribution (Sharda et al., 2018). Finally, the number of observations should be increased to a more representative sample size, as an N of 254 is likely too small.

**6 Conclusion**

Resoundingly, the descriptive analysis has served its purpose of informing the leadership of the retail distribution company about its operational and financial position. Clearly, there is financial variability between inherent to their product\_id, territory\_id, and employee\_id attributes. Also, their predictive capacity has been expanded with a product focus. Moving toward a desired future state of growth is dependent on strategic action and another iteration of this analysis with incremental data.

**References**

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. Journal of Retailing, 93(1), 79-95.

CSU Global. (n.d. a). Module 02: Business Questions. In MIS542: Business Analytics: Spring D 2019 [Class lecture slide]. Retrieved from <https://csuglobal.instructure.com/courses/10900/modules/items/530463>

CSU Global. (n.d. b). Module 08: Portfolio Project. In MIS543: Enterprise Performance Management: Spring B 2020. Retrieved from <https://csuglobal.instructure.com/courses/21042/assignments/428072>

Elliott, A. C., & Woodward, W. A. (2016). SAS essentials: Mastering SAS for data analytics (Second ed.). Hoboken, NJ: Wiley.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (Third ed.).

Marr, B. (2015). Big data: using smart big data, analytics and metrics to make better decisions and improve performance (1st ed.). Retrieved from <https://redshelf.com/book/572147/big-data-572147-9781118965788-bernard-marr>

Sharda, R., Delen, D., & Turban, E. (2018). Business intelligence, analytics, and data science: A managerial perspective (4th ed.).