**Business Analytics with Product Orders Dataset/ SAS Studio**

Didem B. Aykurt

Colorado State University Global

MIS543: Enterprise Performance Management

Dr. John Marlowe

September 24, 2023

## Descriptive Statistics & Hypothesis Test with SAS Studio

SAS Studio is one of the data analysis tools. This is a great tool to create inferential, descriptive, and predictive statistics with just a click without typing code. Let's deep dive into SAS Studio with the products.csv data set. The products data set is a multivariate data set with 30 variables and 951669 rows from the CSU-Global library. Print the variable on the product dataset.

**Figure 1:** Products.csv data set variables list.

Alphabetic List of Variables and Attributes

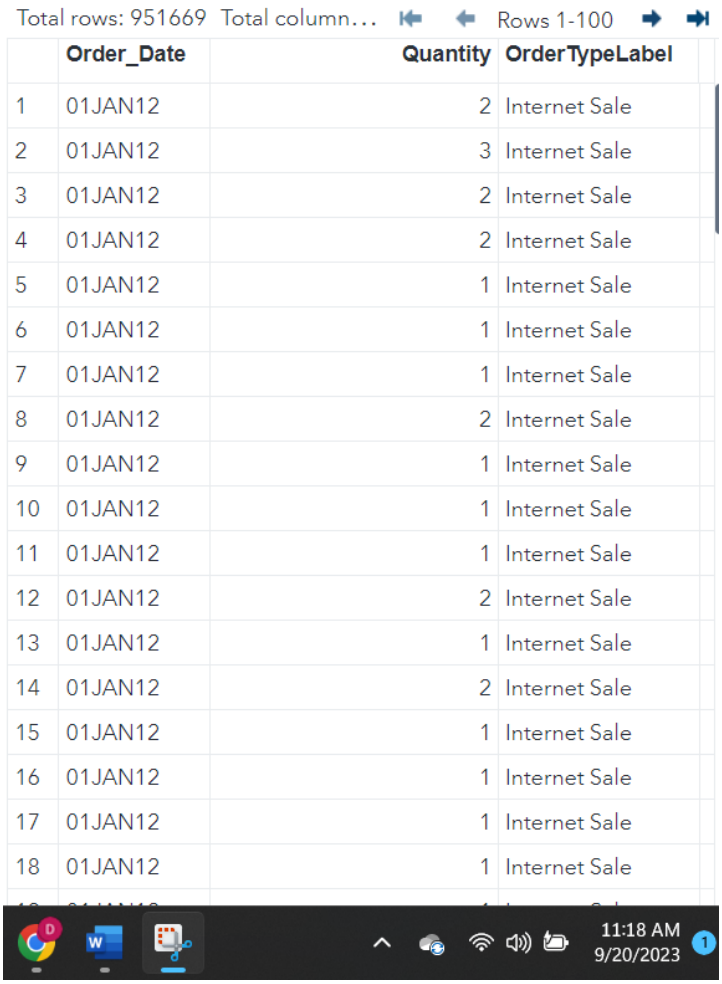
| # | Variable | Type | Len | Format | Informat |
|---|---|---|---|---|---|
| 1 | City | Char | 17 | $17. | $17. |
| 2 | Continent | Char | 13 | $13. | $13. |
| 13 | Cost | Num | 8 | NLNUM12. | NLNUM32. |
| 3 | CountryLabel | Char | 14 | $14. | $14. |
| 25 | Customer_ID | Num | 8 | BEST12. | BEST32. |
| 10 | Delivery_Date | Num | 8 | DATE8. | DATE8. |
| 14 | Discount | Char | 1 | $1. | $1. |
| 26 | Employee_ID | Num | 8 | BEST12. | BEST32. |
| 15 | OrderTypeLabel | Char | 13 | $13. | $13. |
| 9 | Order_Date | Num | 8 | DATE8. | DATE8. |
| 27 | Order_ID | Num | 8 | BEST12. | BEST32. |
| 4 | PostalCode | Char | 7 | $7. | $7. |
| 16 | Product_Category | Char | 24 | $24. | $24. |
| 17 | Product_Group | Char | 21 | $21. | $21. |
| 28 | Product_ID | Num | 8 | BEST12. | BEST32. |
| 18 | Product_Line | Char | 15 | $15. | $15. |
| 19 | Product_Name | Char | 42 | $42. | $42. |
| 11 | Quantity | Num | 8 | BEST12. | BEST32. |
| 12 | RetailPrice | Num | 8 | NLNUM12. | NLNUM32. |
| 5 | State_Province | Char | 22 | $22. | $22. |
| 29 | Street_ID | Num | 8 | BEST12. | BEST32. |
| 6 | Street_Name | Char | 31 | $31. | $31. |
| 20 | SupplierContinent | Char | 13 | $13. | $13. |
| 21 | SupplierCountryLabel | Char | 14 | $14. | $14. |
| 30 | Supplier_ID | Num | 8 | BEST12. | BEST32. |
| 22 | Supplier_Name | Char | 26 | $26. | $26. |
| 7 | xyContinentLat | Num | 8 | BEST12. | BEST32. |
| 8 | xyContinentLon | Num | 8 | BEST12. | BEST32. |
| 23 | xySupContLat | Num | 8 | BEST12. | BEST32. |
| 24 | xySupContLong | Num | 8 | BEST12. | BEST32. |



SAS also has tools to see tables without complex code, such as dragging the dataset from the left side of My Library to drop the coding side and display the table. This way, the user can filter

variables to display. This paper contains a few variables, Order_Date, Quantity, and OrderTypeLabel, that answer two business questions the company needs.

**Figure 2:** List Data task results filtered dataset.



Total rows: 951669 Total column… Rows 1-100

| | Order_Date | Quantity | OrderTypeLabel |
|---|---|---|---|
| 1 | 01JAN12 | 2 | Internet Sale |
| 2 | 01JAN12 | 3 | Internet Sale |
| 3 | 01JAN12 | 2 | Internet Sale |
| 4 | 01JAN12 | 2 | Internet Sale |
| 5 | 01JAN12 | 1 | Internet Sale |
| 6 | 01JAN12 | 1 | Internet Sale |
| 7 | 01JAN12 | 1 | Internet Sale |
| 8 | 01JAN12 | 2 | Internet Sale |
| 9 | 01JAN12 | 1 | Internet Sale |
| 10 | 01JAN12 | 1 | Internet Sale |
| 11 | 01JAN12 | 1 | Internet Sale |
| 12 | 01JAN12 | 2 | Internet Sale |
| 13 | 01JAN12 | 1 | Internet Sale |
| 14 | 01JAN12 | 2 | Internet Sale |
| 15 | 01JAN12 | 1 | Internet Sale |
| 16 | 01JAN12 | 1 | Internet Sale |
| 17 | 01JAN12 | 1 | Internet Sale |
| 18 | 01JAN12 | 1 | Internet Sale |



The company seeks to understand the customer experience and convenience of ordering types like Internet or Retail Sales. By asking the below questions, the company can gain insights into how these order types differ in terms of the overall shopping experience, ease of use, and accessibility.

1. What are the differences between Internet and retail sales regarding customer experience and convenience?

2. What are the most order quantities of order type are Internet and Retail sales?

Next, create the null and alternative hypotheses for each business and a hypothesis test that could be conducted to compare internet sales and retail sales in terms of the highest percentage of order type range: the question as follows:

- The first business problem hypotheses are as follows:
    - **Null hypothesis:** There is no significant difference between the customer experience and convenience of Internet and retail sales.
    - **Alternative Hypothesis:** A significant difference exists between the customer experience and convenience of Internet and retail sales.
- The second business problem hypotheses are as follows:
    - **Null Hypothesis:** The highest quantity of order type range for internet sales and retail sales is the same.
    - **Alternative Hypothesis:** The highest quantity of order type range for internet sales and retail sales is different.

Before conducting a chi-square test, it is generally recommended to perform descriptive statistics on the data. Descriptive statistics summarize and organize the characteristics of a data set, providing valuable insights into the distribution and central tendencies of the variables. Some standard descriptive statistics that can be calculated before a chi-square test include: Frequency Distribution: A frequency distribution table displays the number or percentage of observations in each category of a categorical variable. This helps understand the distribution of responses and identify any imbalances or patterns.

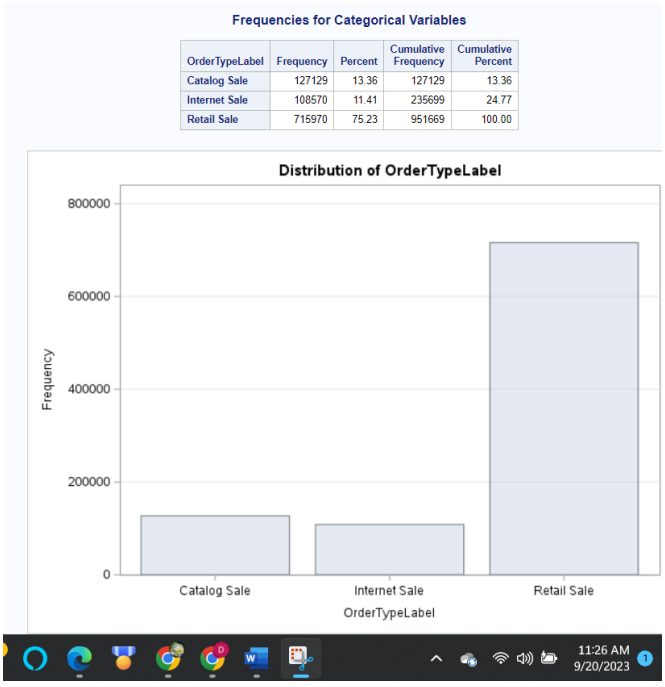
Measures of Central Tendency: The mean, median, and mode can provide insights into a variable's average or typical value. These measures help understand the central tendencies of the data.

Measures of Variability: The range, variance, and standard deviation quantify the spread or dispersion of the data. They provide information about how much the values deviate from the central tendency.

**Frequency Distribution**

Under the Data task list has a Characterize Data help to manipulate data needed and explore. Selected the Order_Date, Quantity, and OrderTypeLabel from the Data tab. Then, the Options tab has a select option for categorical and numerical variables. The result is shown in Figure 2.

**Figure 3:** The result window of the Characterize Data task by OrderTypeLabel variable.




The product dataset has three order types: Internet, Retail, and Catalog Sales. The frequency column represents the number of occurrences for each order type as the order highest to

lowest order type is Retail Sales 715970, Catalog Sales 127129, and Internet Sales 108570. The table allows us to see the percentage of each order type as Retail Sales at 75.23%. The frequency distribution histogram summarizes the distribution of order types and the number of occurrences for each type.

**Measures of Central Tendency**

**Figure 4:** Result of Summary Statistic task with OrderTypeLabel by Quantity.

| Analysis Variable : Quantity | | | | | | | |
|---|---|---|---|---|---|---|---|
| OrderTypeLabel | N Obs | Mean | Std Dev | Median | Variance | Mode | Range |
| Catalog Sale | 127129 | 1.7477916 | 0.9157979 | 2.0000000 | 0.8386858 | 1.0000000 | 8.0000000 |
| Internet Sale | 108570 | 1.7545270 | 0.9216709 | 2.0000000 | 0.8494773 | 1.0000000 | 9.0000000 |
| Retail Sale | 715970 | 1.6545847 | 0.8924216 | 1.0000000 | 0.7964163 | 1.0000000 | 9.0000000 |


**Mean:** The mean is calculated by summing up all the Quantity values and dividing by the total number of values. For the dataset above, the mean number of orders for Internet Sales is 1.75, and for Retail Sales, it is 1.65, which shows that the Retail and Internet order average quantity per order number is close but not the same.

**Median:** The median is the middle value in an ordered dataset. To find the median, sort the values in ascending order and select the central value. If there is an even number of values, take the average of the two middle values. In this case, Internet Sales and Retail Sales have an even number, so the median number of orders for both types is 2 and 1; half the data value falls below the median, and half the data value falls above the median. As a result, the median tells half of the Internet Sales order quantity 2, and the Retail Sales order quantity is more than half of the data point 1. Buyers from the internet mostly buy two products, and customers from

the retailer buy one. The mean is more affected by outliers than the median. This tells Internet Sales median and mean so close that might normal distribution that result helps to reject or not reject the hypothesis with a chi-square result. The Retail Sales' median number is smaller than the mean, which might skew.

**Measures of Variability**

**Figure 5:** Mosaic plot shows order type and quantity data.

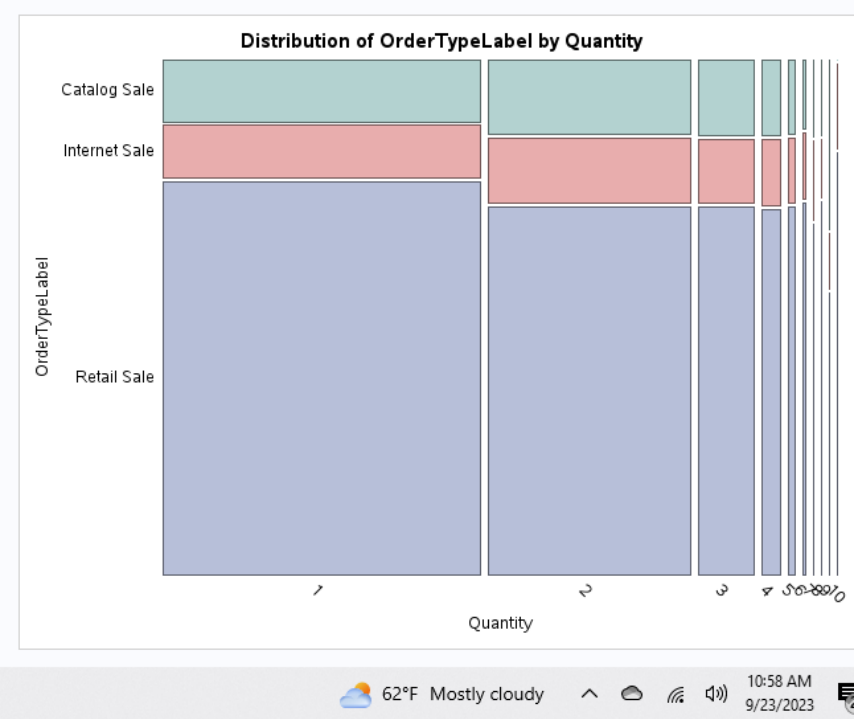
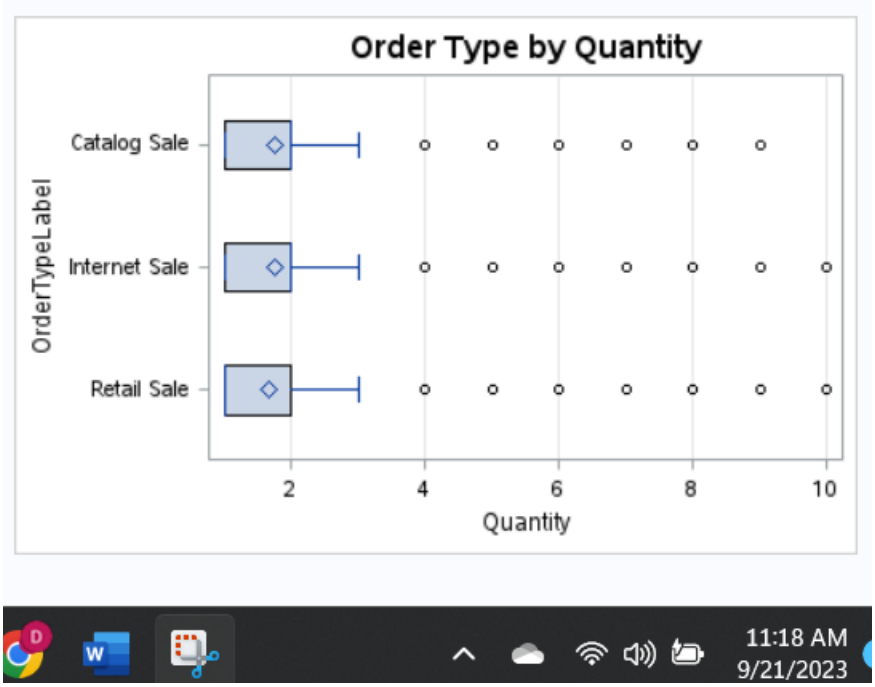


**Figure 6:** Box plot represents order type and quantity.




A box plot, also known as a box and whisker plot, is a graphical representation of the distribution of numerical data. It displays the five-number summary of a dataset, which includes the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The box represents the interquartile range (IQR), which is the range between the first quartile (Q1), which means the 25$^{th}$ percentile, and the third quartile (Q3), which represents the 75$^{th}$ percentile. 50% of all data values and the distance between the first and second quartile. The box plot contains lines on the left side and right side of the box showing all the data points within 1.5 interquartile ranges below the first quartile or above the third quartile. The small circles also represent outlier values with more than 1.5 interquartile ranges. The box plot shows the data points outside the whiskers are considered outliers. All the order types of boxes represent the interquartile range below quantity 2. Also, the mosaic plot shows the most quantity order 1 and

2. As a result, it is mostly per order type quantity 1 or 2. The others are outliers 4,6,8 and 10 for Internet Sales and Retail Sales.

A box plot provides a concise summary of the data distribution, allowing you to identify the dataset's central tendency, spread, and skewness. It is beneficial for comparing multiple data groups and identifying differences in their distributions.

**Chi-Square Statistics**

The chi-square test is a statistical hypothesis test used to analyze categorical data and determine if there is a significant association between two categorical variables. It compares the observed frequencies of different categories with the expected frequencies to assess whether the differences are statistically significant. There are two main types of chi-square tests:

1- Chi-Square Goodness of Fit Test: This test compares the observed frequency distribution of a single categorical variable with the expected frequency distribution. It determines whether the observed data significantly deviates from the expected distribution.

2- Chi-Square Test of Independence: This test examines the relationship between two categorical variables and determines if they are independent or related. It compares the observed frequencies in a contingency table with the expected frequencies under the independence assumption.

In this case, the Table Analysis task contains a chi-square test to see the relationships between variables. The column is OrderTypeLabel, and the row variable is Quantity. The OPTIONS tab has many options for this time; the figure below shows all options for this test.

**Figure 7:** OPTION tab for a two-way table

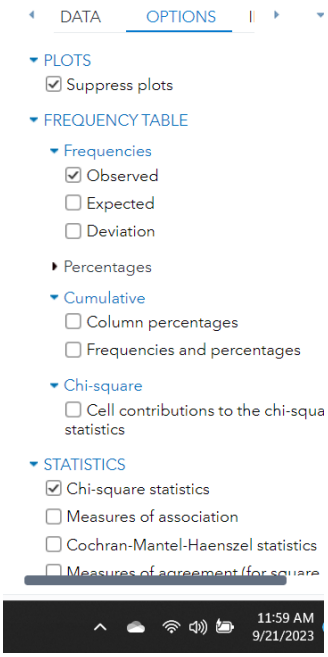
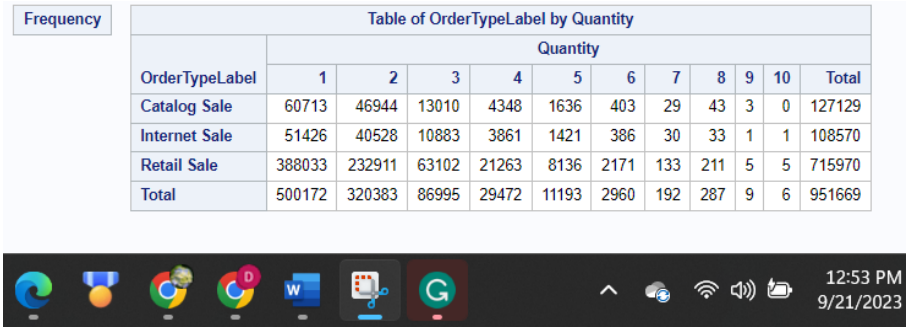



**Figure 8:** Result of Order Type by Quantity.



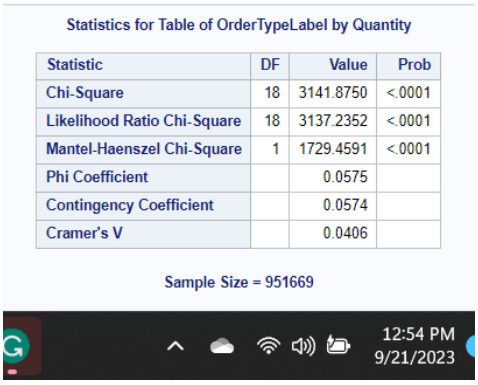
Frequency

Table of OrderTypeLabel by Quantity

| | Quantity | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| OrderTypeLabel | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
| Catalog Sale | 60713 | 46944 | 13010 | 4348 | 1636 | 403 | 29 | 43 | 3 | 0 | 127129 |
| Internet Sale | 51426 | 40528 | 10883 | 3861 | 1421 | 386 | 30 | 33 | 1 | 1 | 108570 |
| Retail Sale | 388033 | 232911 | 63102 | 21263 | 8136 | 2171 | 133 | 211 | 5 | 5 | 715970 |
| Total | 500172 | 320383 | 86995 | 29472 | 11193 | 2960 | 192 | 287 | 9 | 6 | 951669 |



**Figure 9:** Result of Chi-square statistics.



Statistics for Table of OrderTypeLabel by Quantity

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 18 | 3141.8750 | <.0001 |
| Likelihood Ratio Chi-Square | 18 | 3137.2352 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 1729.4591 | <.0001 |
| Phi Coefficient | | 0.0575 | |
| Contingency Coefficient | | 0.0574 | |
| Cramer's V | | 0.0406 | |

Sample Size = 951669

To perform a chi-square test, a variable must have categorical data and construct a contingency table that summarizes the observed frequencies for each combination of categories. The test calculates a test statistic called the chi-square statistic ($\chi^2$) and compares it to the critical value from the chi-square distribution to determine statistical significance. Suppose the chi-square test yields three p-values below a predetermined significance level (e.g., 0.05), so the result of the chi-square test represents a p-value .0001 smaller than .05. In that case, it suggests that the observed frequencies significantly differ from the expected frequencies, indicating a relationship between the variables. The three p-values are below the significance level; it provides evidence to reject the two null hypotheses and concludes that a statistically significant association exists between the order types and quantity range. The frequency tables support all the order types, and the quantity ranges differ.

**Conclusion**

The Retail Sales percentage of 75.23% is more significant than other order types. Per order, the highest number range is 1 to 2. Thus, the company might focus on the retail business to improve customer service and sales of products to satisfy customer volume or the next step to find why Internet Sales are less than other order types. Before starting this case, most companies would expect Internet Sales more than different order types because Internet sales offer the convenience of product selection, accessibility, and doorstep delivery, saving customers time and effort. Customers can have products delivered directly to their homes or preferred locations. Retail sales require customers to travel to a physical store, which may involve additional time and transportation costs.

Might focus on Retail sales to provide the opportunity for face-to-face interaction with sales associates, who can offer personalized assistance, recommendations, and immediate answers to customer queries. This personal touch may enhance the overall customer experience. In contrast, internet sales rely on digital interfaces and may lack the same level of personal interaction.

**References**

Cody, R.(2021). *A Gently Introduction to Statistics Using SAS Studio in the Cloud*. SAS Institute.

ISBN: 9781954844476.

Turney, S. (2023). *Chi-Square ($X^2$) Tests, Types, Formula & Examples*. https://www.scribbr.com/statistics/chi-square-tests/

Mcleod, S. (2023). *Box Plot Explained: Interpretation, Examples, & Comparison*. https://www.simplypsychology.org/boxplots.html

Bhandari, P. (2020). *Descriptive Statistics | Definitions, Types, Examples.* *https://www.scribbr.com/statistics/descriptive-statistics/*