Data Manipulation of OrderNorthwind Data/ SAS Studio

Didem Bulut Aykurt

MIS540-1 – Introduction to Business Intelligence

Colorado State University-Global Campus

Dr. Alin Tomoiaga

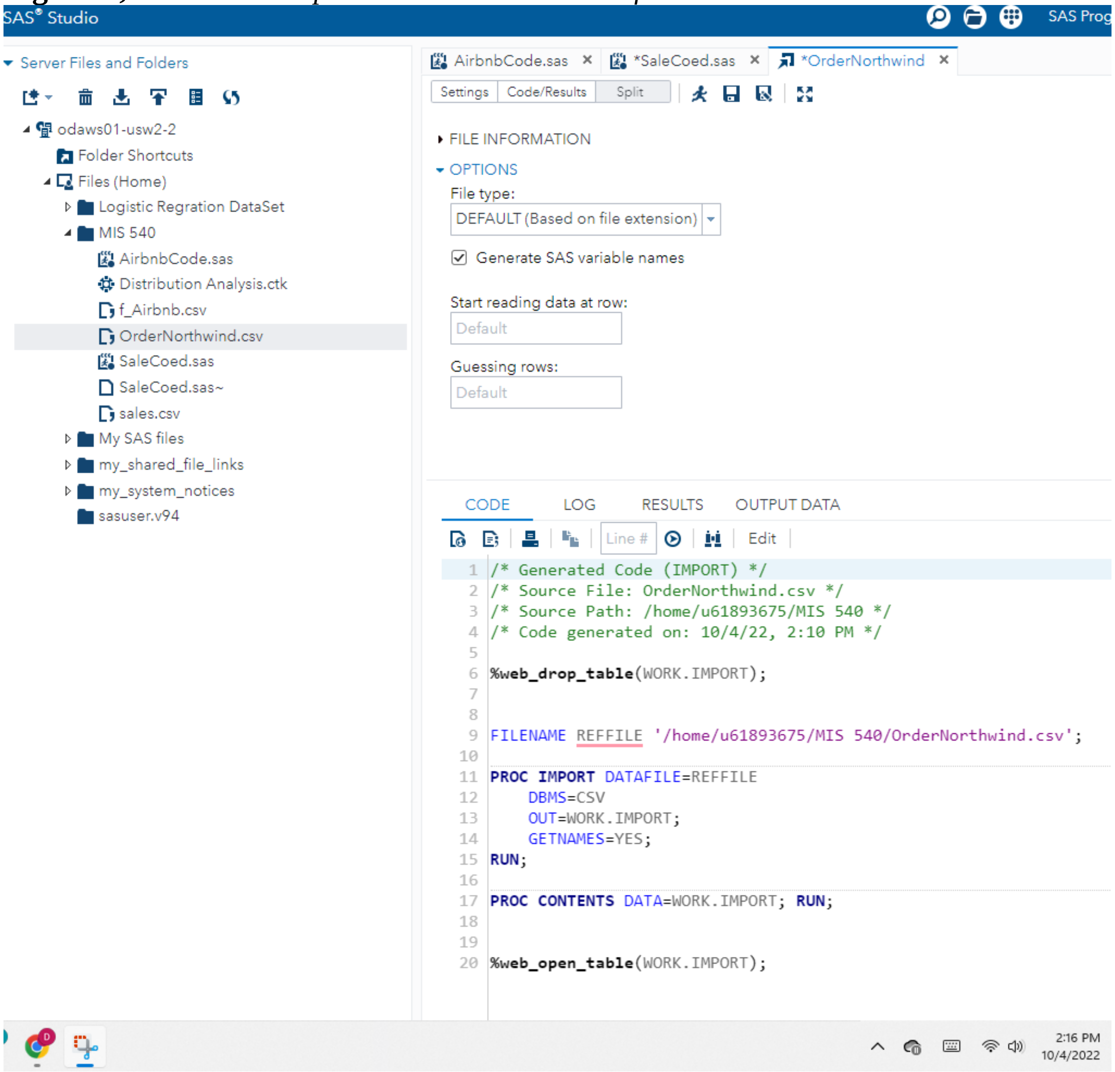
October 7, 2022

Frequency Distribution

A frequency analyst is a statistical tool that helps to find frequency per categorical variables. A frequency distribution count totals the numbers in each category based on how many times it appears. I will use the OrderNorthwind dataset to create a frequency analyst to answer business questions.

The first step is uploading and importing the OrderNorthwind CSV file into SAS Studio.

Figure 1; *SAS code to import OrderNorthwind CSV file into SAS Studio*



File type: DEFAULT (Based on file extension)

☑ Generate SAS variable names

Start reading data at row: Default

Guessing rows: Default

```
/* Generated Code (IMPORT) */
/* Source File: OrderNorthwind.csv */
/* Source Path: /home/u61893675/MIS 540 */
/* Code generated on: 10/4/22, 2:10 PM */

%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/home/u61893675/MIS 540/OrderNorthwind.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.IMPORT;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;


%web_open_table(WORK.IMPORT);
```

Figure 2; *SAS Output Data of the OrderNorthwind dataset.*

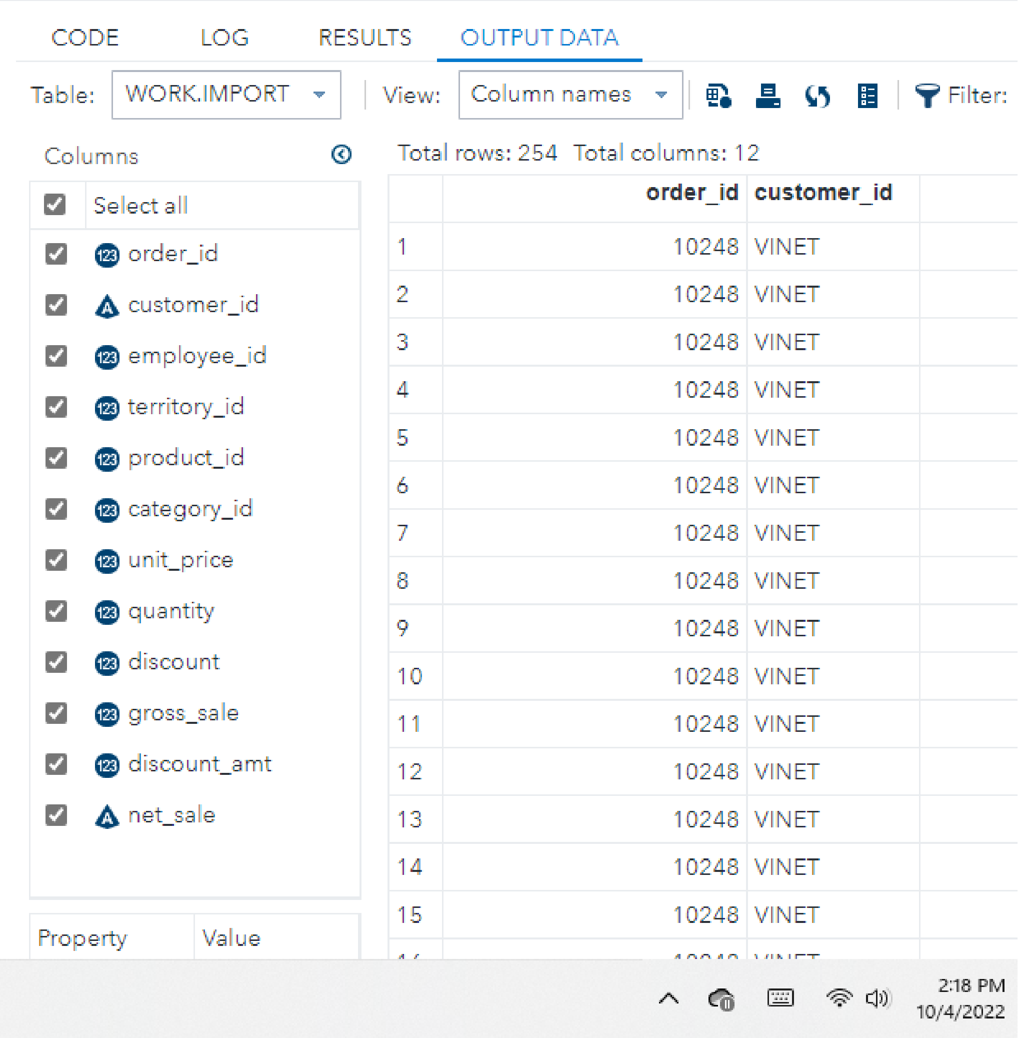


Table: WORK.IMPORT View: Column names

Columns

☑ Select all
☑ order_id
☑ customer_id
☑ employee_id
☑ territory_id
☑ product_id
☑ category_id
☑ unit_price
☑ quantity
☑ discount
☑ gross_sale
☑ discount_amt
☑ net_sale

Property Value

Total rows: 254 Total columns: 12

	order_id	customer_id
1	10248	VINET
2	10248	VINET
3	10248	VINET
4	10248	VINET
5	10248	VINET
6	10248	VINET
7	10248	VINET
8	10248	VINET
9	10248	VINET
10	10248	VINET
11	10248	VINET
12	10248	VINET
13	10248	VINET
14	10248	VINET
15	10248	VINET

The OrderNorthwind dataset contains 254 observations with 12 attributes as six continuous variables unit_price, quantity, discount, gross_sale, discount_atm, and net_sale. The other six categorical variables are order_id, customer_id, employee_id, territory_id, product_id, and category_id. In this assignment, I utilized the OrderNorthwind dataset variables' product_id, category_id, unit_price, and gross_sale. Those variables help to answer business questions. These are the business problems that will explain how to improve business needs as:

1. How does unit price affect the gross sale? Is there a positive relationship between the two variables?

2. Is there a difference in the category type based on the gross sale? Which category of product's gross sale is most valuable?

Then we need to create the null hypothesis and alternative hypotheses for each business question as follows:

- First business problem hypotheses:

  Null hypotheses: It is no relation between the unit price and gross sale.

  Alternative Hypothesis: They have a relation between two variables.

  We reject the null hypothesis because the p-value is lower than 0.05 in figure 4 and figure 6 table results. Figure 6 scatterplot has a correlation result of 61.6% positive relation. The unit price average is accurate for increasing gross sales, but the gross sale average is 492.97 lower than expected than the unit price average. Both variables have positive skew and right-sight outliers. That also shows the relation between the two variables.

- Second business problem hypothesis:

  Null hypothesis: There is no difference between all gross sales(gross_sale) of product category types(categoty_id).

  Alternative hypothesis: at least one group significantly differs from the overall gross sale(gross_sale) of the product's category(category_id).

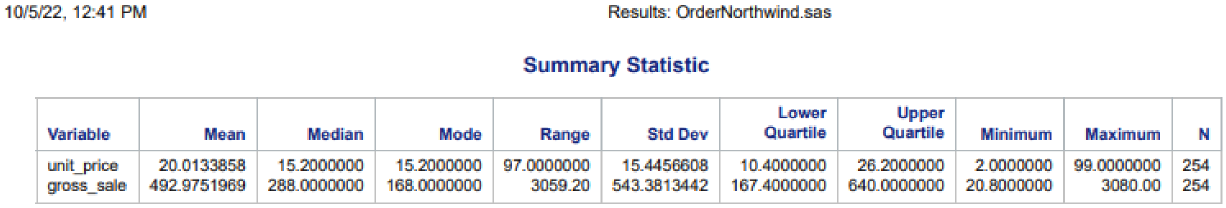
  Figure 7 table and histogram chart show that the total gross sales of the category are different. Figure 8 has a one-way ANOVA test result showing a p-value lower than 0.05 and an F value higher than 1. All the category means and std dev results are different from each other. Thus, reject the null hypothesis.

Now, both business questions have a hypothesis, and the analyst for all the ideas is a descriptive and frequency analyst because one variable is categorical and the other is numerical.

Descriptive Statistic

Let’s look at the numerical variable’s unit_price and gross_sale as summary statistics help to see a central tendency: where are the most data point and data points of the spread that is the distribution.

Figure 3: *The result of a summary statistic for “unit_price” and “gross_sales” in SAS Studio.*



10/5/22, 12:41 PM Results: OrderNorthwind.sas

Summary Statistic

Variable	Mean	Median	Mode	Range	Std Dev	Lower Quartile	Upper Quartile	Minimum	Maximum	N
unit_price	20.0133858	15.2000000	15.2000000	97.0000000	15.4456608	10.4000000	26.2000000	2.0000000	99.0000000	254
gross_sale	492.9751969	288.0000000	168.0000000	3059.20	543.3813442	167.4000000	640.0000000	20.8000000	3080.00	254

Mean shows the average unit price on our data set average of unit_price 20.01 and gross_sale 492.98. As we can say product price average is accurate, but the total gross sale average is lower than the product price average is correct.

The **Median** tells the middle value of the unit price and gross sale as half-unit price data points lower than 15.20 and higher than that. The gross sale middle value is 288, half the data point lower than an upper than 288.

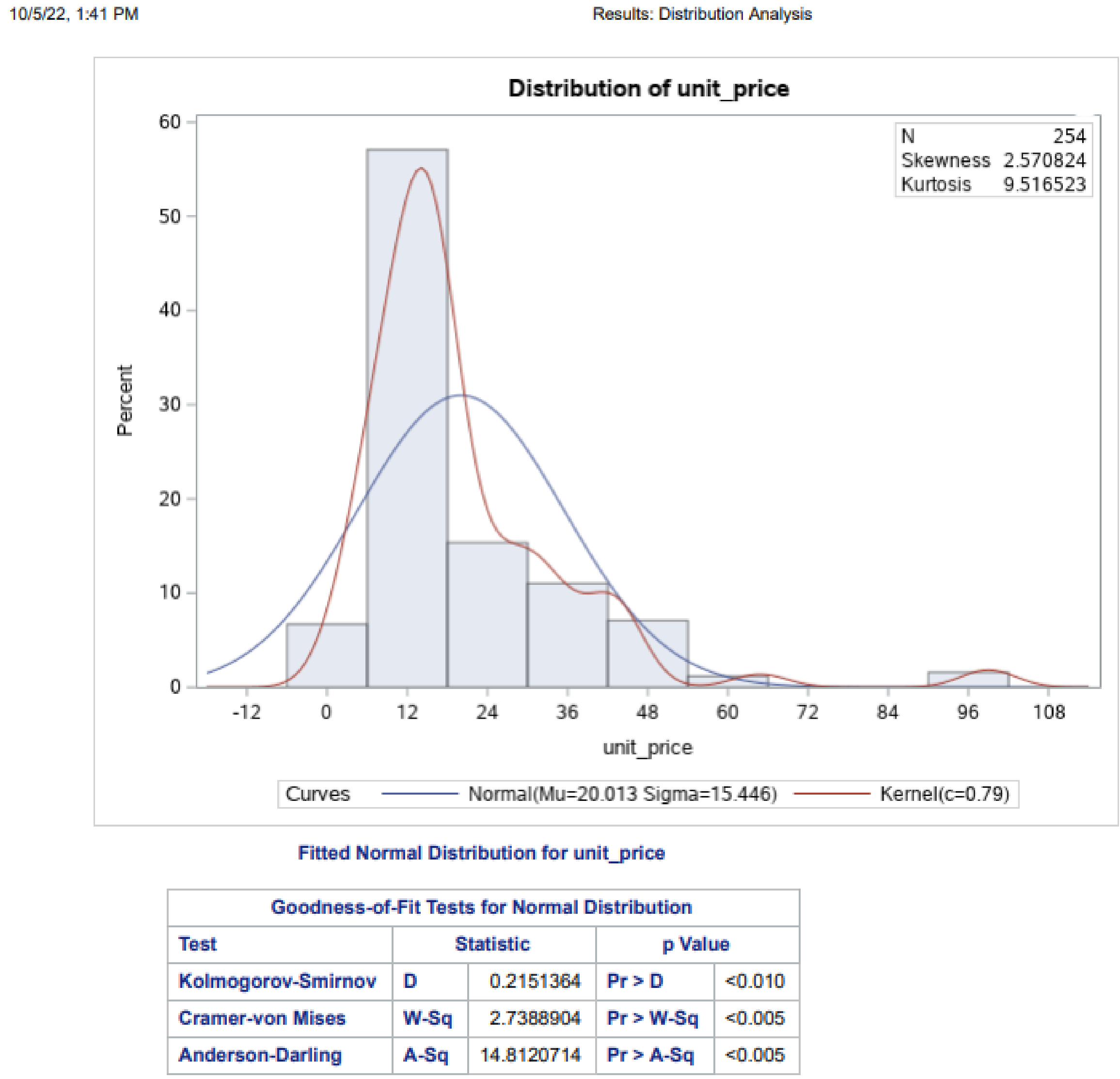
Mode helps to find one number repeating more than one that unit price median number equals the mode number 15.20, and the gross sale mode point is 168. if the unit price’s mean number equals the mode and median, we can say normal distribution as the skewness result should show how far from a normal distribution.

Q1/Lower Quartile shows 25% of unit price data split above data point 10.4, and gross sale first quartile is 167.40.

Q2/Upper Quartile helps to see 75% of all unit price data point fall below the third quartile 26.20. And the gross sale third quartile point is 640, which shows all the data points stay halfway between the highest and medial numbers.

Distribution Analysis

Figure 4: *The resulting distribution analysis is a histogram of the 'unit_price' variable on SAS Studio.*



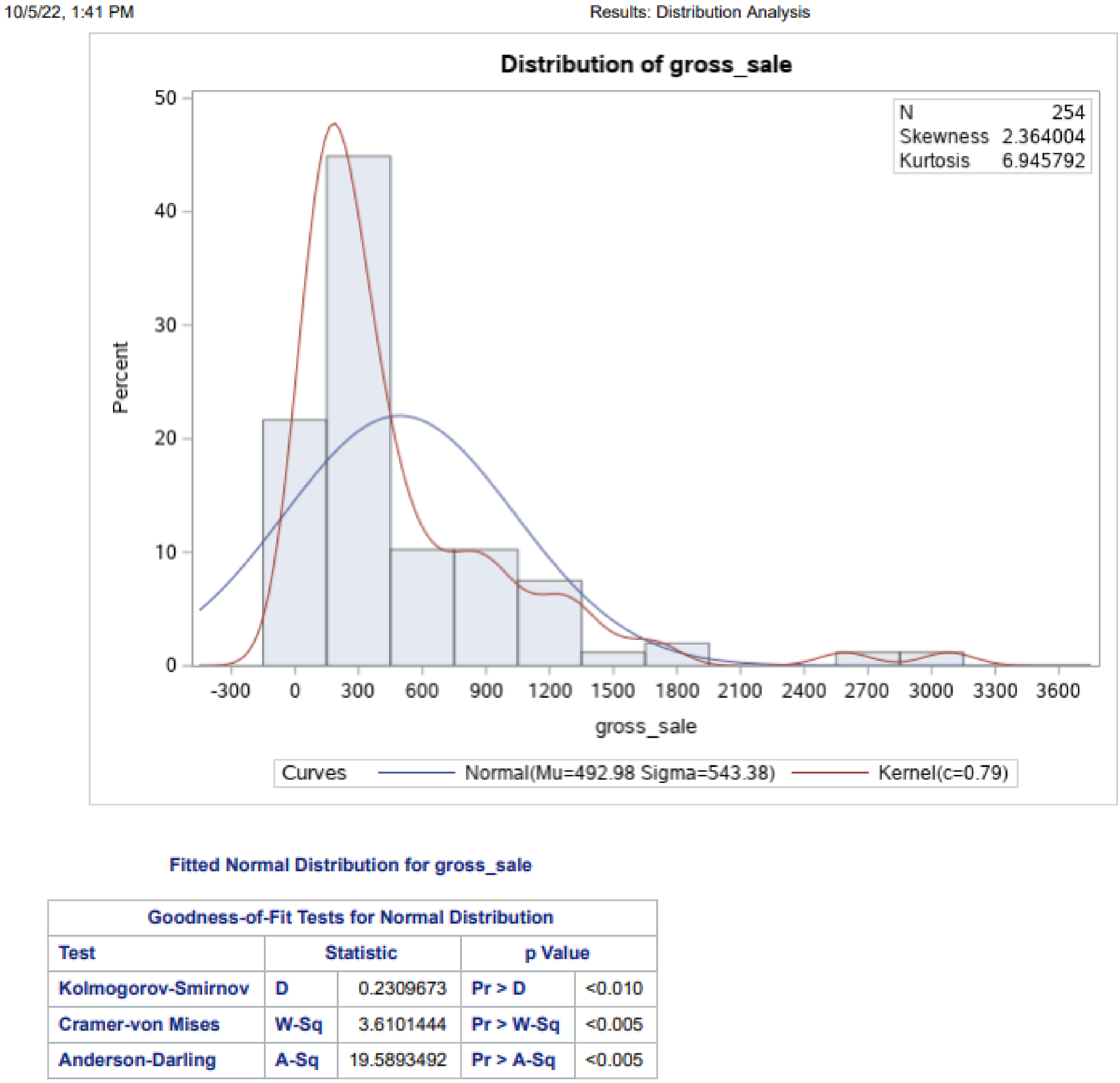
10/5/22, 1:41 PM Results: Distribution Analysis

Fitted Normal Distribution for unit_price

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.2151364	Pr > D	<0.010
Cramer-von Mises	W-Sq	2.7388904	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	14.8120714	Pr > A-Sq	<0.005

In summary statistics, we mention unit price normal distribution as figure 4 has a **skewness** of price unit 2.57 has positive skew or right skew. **Kurtosis** tells outliers; unit price kurtosis higher than zero means leptokurtic says the unit price has the correct side outlier. The goodness of fit test shows all p-value lower than 0.05, which means the unit price dataset is statistically significant or can say sample data fit population data.

Figure 5: *The resulting distribution analysis is a histogram of the 'gross_sale' variable on SAS Studio.*



10/5/22, 1:41 PM

Results: Distribution Analysis

Fitted Normal Distribution for gross_sale

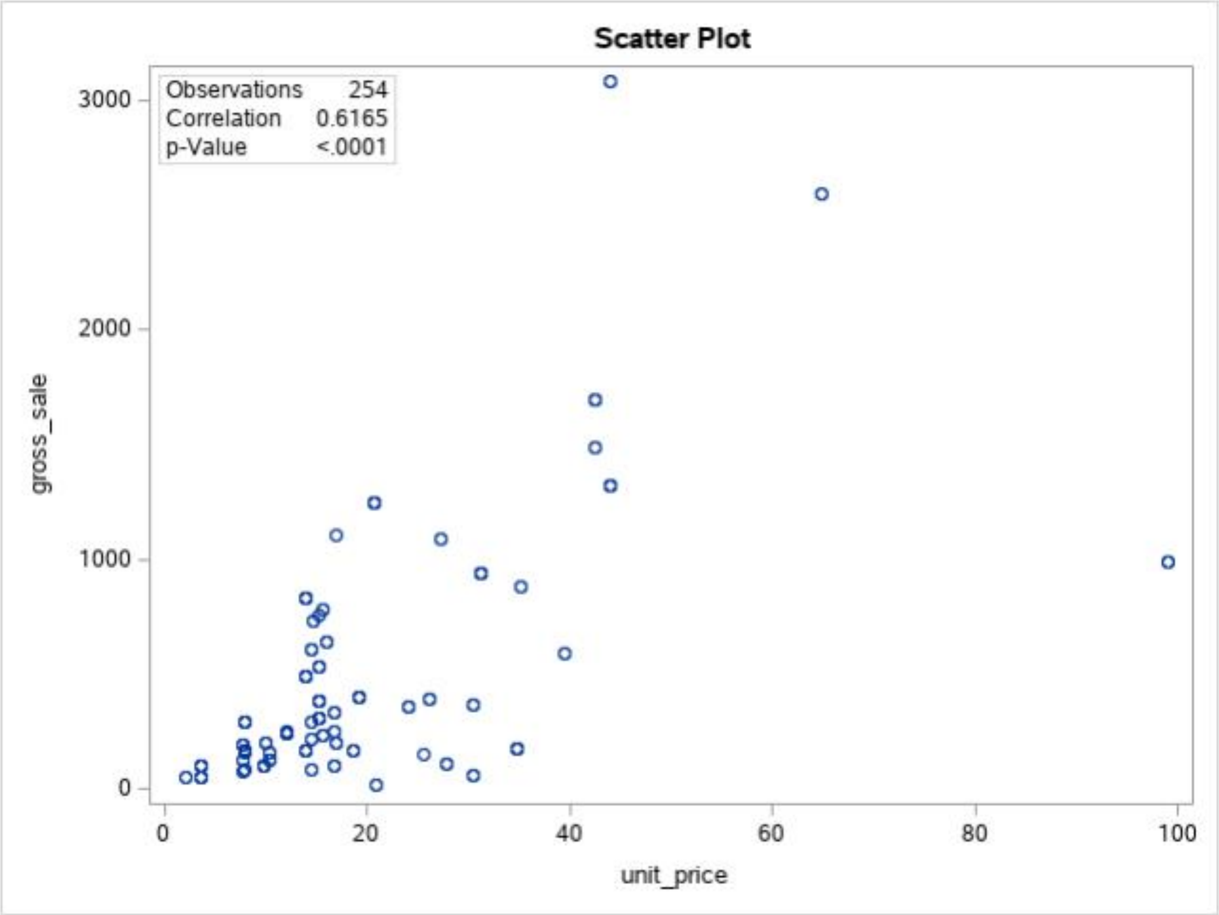
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.2309673	Pr > D	<0.010
Cramer-von Mises	W-Sq	3.6101444	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	19.5893492	Pr > A-Sq	<0.005

Figure 5 shows a gross sale dataset distribution result; skewness is 2.36 positive, and gross sale has a positive or right skew. **Kurtosis** of gross sale kurtosis is leptokurtic 6.94 higher than zero, meaning the gross sale has a right-side outlier.

The goodness of fit test results in three of a p-value lower than 0.05, which means the gross sale dataset is significantly statistical as we can use this dataset as a sample for population results.

Correlation Analysis

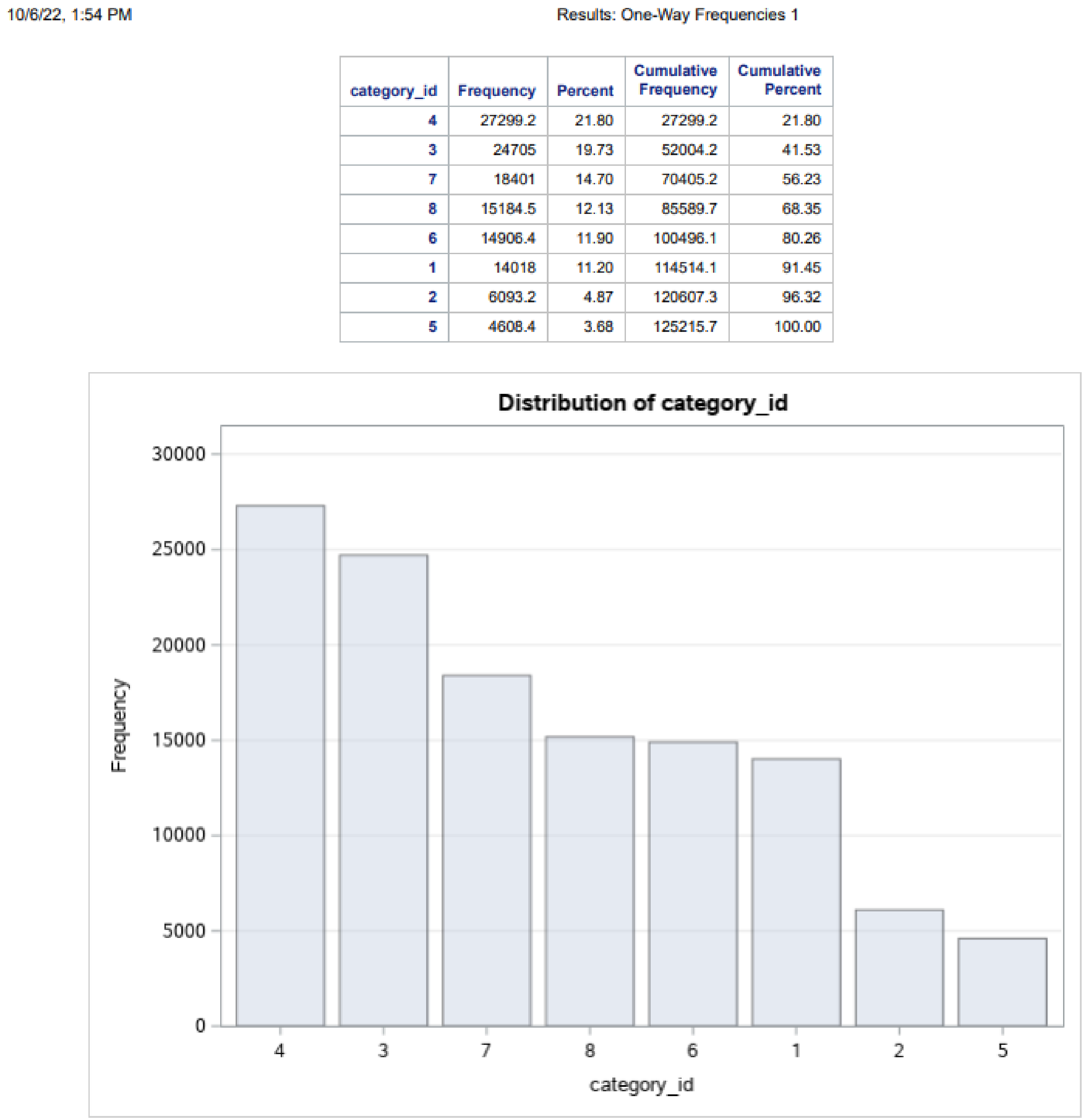
Figure 6: *The result of correlation analysis for the independent variable "unit_price" and dependent variable "gross_sale" in SAS Studio.*



The unit price and gross sale have a positive correlation of 61%; if the correlation results are higher than 50%, that is a strong positive correlation. P-value is lower than 0.05. As a result, it is statistically significant.

One-Way Frequencies

Figure 7: *The frequency analysis result for "category_id" and "gross_sale" in SAS Studio.*



10/6/22, 1:54 PM Results: One-Way Frequencies 1

category_id	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	27299.2	21.80	27299.2	21.80
3	24705	19.73	52004.2	41.53
7	18401	14.70	70405.2	56.23
8	15184.5	12.13	85589.7	68.35
6	14906.4	11.90	100496.1	80.26
1	14018	11.20	114514.1	91.45
2	6093.2	4.87	120607.3	96.32
5	4608.4	3.68	125215.7	100.00

There are eight categories for product sales on the one-way frequency analyst table. The **Frequency** column shows how many data points fell into the product category. The **Percent** column specifies the percentage of data points in that category. The **Cumulative Frequency** column indicates the addition of all the numbers in the Frequency column above and includes the current row. The last column on the table is the **Cumulative Percent,** which shows the addition of all the Percent columns up to the current row. The OrderNorthwind dataset doesn't have a missing value.

Category id 4 has the highest gross sales, 27299.2 frequency of 21.80% of total gross sales from category 4.

Then category id is three higher frequency 24705, which means 19.73% of total gross sales derived from category 3.

Next, category 7 has an 18401 frequency of 14.70% of total gross sales obtained from Category 7.

Category 8's frequency is 15184.5, 12.13% of total gross sales acquired from category 8.

Category 6 shows frequency number 14906.4, 11.90% of total gross sales from category 6.

Category 1 has a low-frequency number of 14018, 11.20% of total gross sales derived from Category 1.

The last two lower percent and frequency category ID sequences of numbers 2 and 5 had cumulative frequencies of 10701.6, both categories' cumulative percentages of 8.55%.

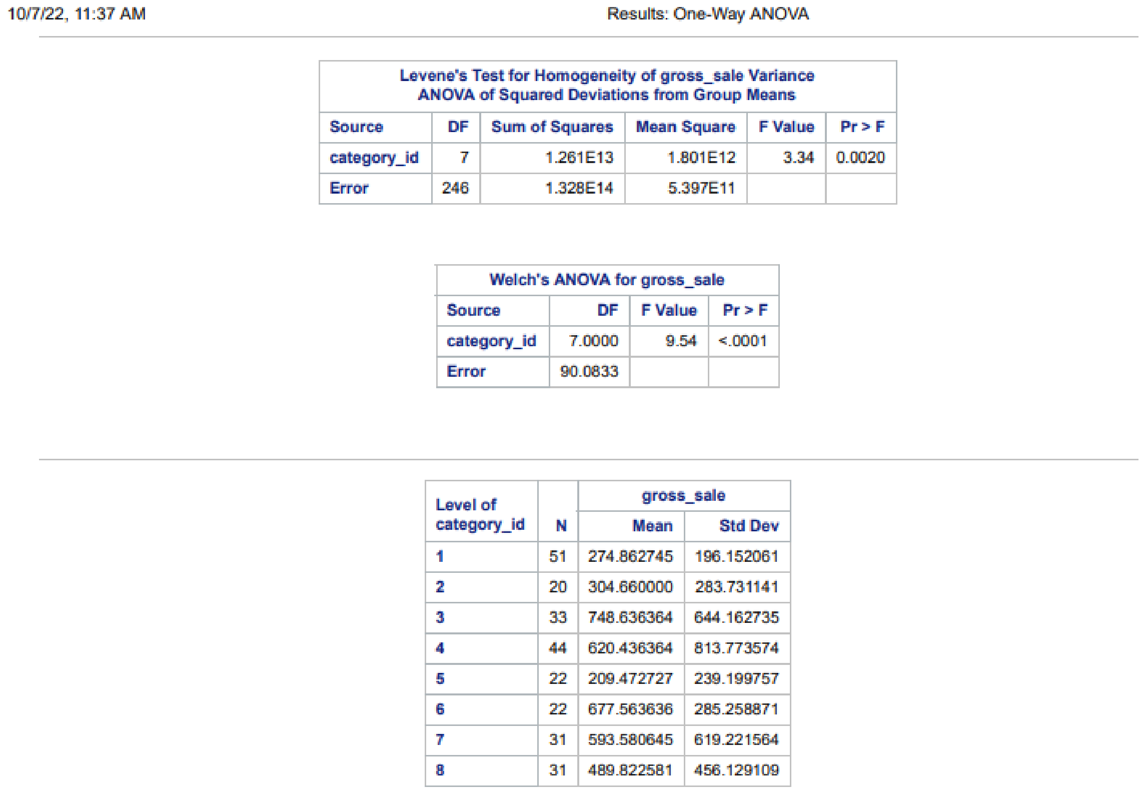
One-Way ANOVA

ANOVA stands for Analysis of Variance, a statistical method to compare the means of two or more groups. It can help you test whether there is a significant difference among the groups or if the observed variation is due to random chance.

There are different types of ANOVA tests, depending on the number and nature of the independent variables (the factors you manipulate) and the dependent variable (the outcome you measure). Some common types are:

One-way ANOVA: This is used when you have one independent variable with two or more levels (groups) and one dependent variable. For example, you can use a one-way ANOVA to compare the mean test scores of gross sales from three different categorical varaible.

Figure 8: The result of the one-way ANOVA test for the dependent variable “gross_sale” and “category variable “category_id.”



10/7/22, 11:37 AM Results: One-Way ANOVA

Levene's Test for Homogeneity of gross_sale Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
category_id	7	1.261E13	1.801E12	3.34	0.0020
Error	246	1.328E14	5.397E11		

Welch's ANOVA for gross_sale

Source	DF	F Value	Pr > F
category_id	7.0000	9.54	<.0001
Error	90.0833		

Level of category_id	N	gross_sale	
		Mean	Std Dev
1	51	274.862745	196.152061
2	20	304.660000	283.731141
3	33	748.636364	644.162735
4	44	620.436364	813.773574
5	22	209.472727	239.199757
6	22	677.563636	285.258871
7	31	593.580645	619.221564
8	31	489.822581	456.129109

The ANOVA for analysis of variance test helps to compare the means with one dependent and one independent categorical variable with a statistical result like F value, p-values, mean, and Std Dev. As the table shows, a p-value lower than 0.05 and an F value more significant than 1. Also, all category mean and std Dev is different. Thus, we can reject the null hypothesis.

Conclusion

Reference,

Sharda, R., Delen, D., & Turban, E. (2022). Business intelligence, analytics, and data science: A managerial perspective (4th ed.). Pearson.

Elliott, A. C., & Woodward, W. A. (2023). SAS Essentials: Mastering SAS for data analytics (3rd ed.). John Wiley & Sons.

KentStateUnivesty.edu. SAS Tutorials: Frequency Tables Using Proc Freq.

The OrderNorthwind.csv dataset from CSU-Global University.

https://libguides.library.kent.edu/SAS/Frequencies#:~:text=SAS%20normally%20orders%20the%20rows%20of%20the%20frequency,PROC%20FREQ%20DATA%3Dsample%20ORDER%3Dfreq%3B%20TABLE%20State%20Rank%3B%20RUN%3B