Descriptive Statistic with Products Dataset/ SAS Studio

Didem B. Aykurt

Colorado State University Global

MIS543: Enterprise Performance Management

Dr. John Marlowe

September 17, 2023

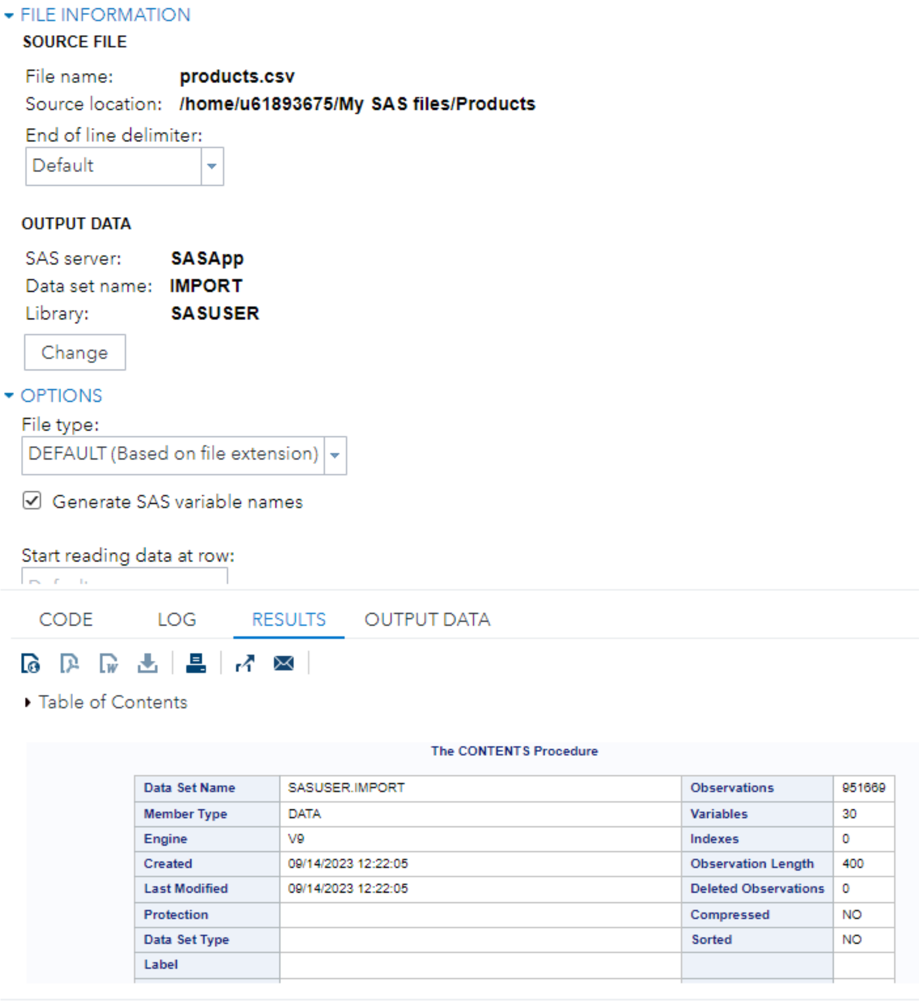
Descriptive Analytics

Descriptive analytics is the essential step of data analysis. This stage has a few critical topics; the summary measures of central tendency include the mean, median, and mode and historical data to understand better standard deviation, variance, range, and the kurtosis and skewness and prepare the data for predictive analytics. For example, if a variable is highly skewed, the variable may need to be normalized to produce a more accurate model. Descriptive statistics has three types of analysis techniques: distribution helps to find the frequency of each value, central tendency analyzes the averages of the value, and variability examines how data points spread out the deals.

Additionally, this project will touch on Summary Statistics and One-Way Frequencies to develop and prepare the dataset for predictive models. Discuss the main topics with the Products Dataset's summary statistics in the Task and Utilities drop-down list results Output Window. Before starting, as descriptive analysts, we should import the products.csv file into the SASUSER library. If the user comes from school or studies, SASUSER, just for reading, should create a folder named Products and upload the products.csv file into the Products folder. Then, import the products.csv file into the SASUSER library, as Figure 1 shows the output.

The products' data sets contain retail company sales, customers, employees, orders, and product information. I used the products.csv dataset, which includes five years of product order data and has 30 variables and 951669 observations. Figure 2 shows each variable name, data type, and format.

Figure 1: Import the products.csv file into the SASUSER library from the Products folder in SAS Studio.



SOURCE FILE

File name: **products.csv**
Source location: **/home/u61893675/My SAS files/Products**
End of line delimiter: Default

OUTPUT DATA

SAS server: **SASApp**
Data set name: **IMPORT**
Library: **SASUSER**

File type: DEFAULT (Based on file extension)

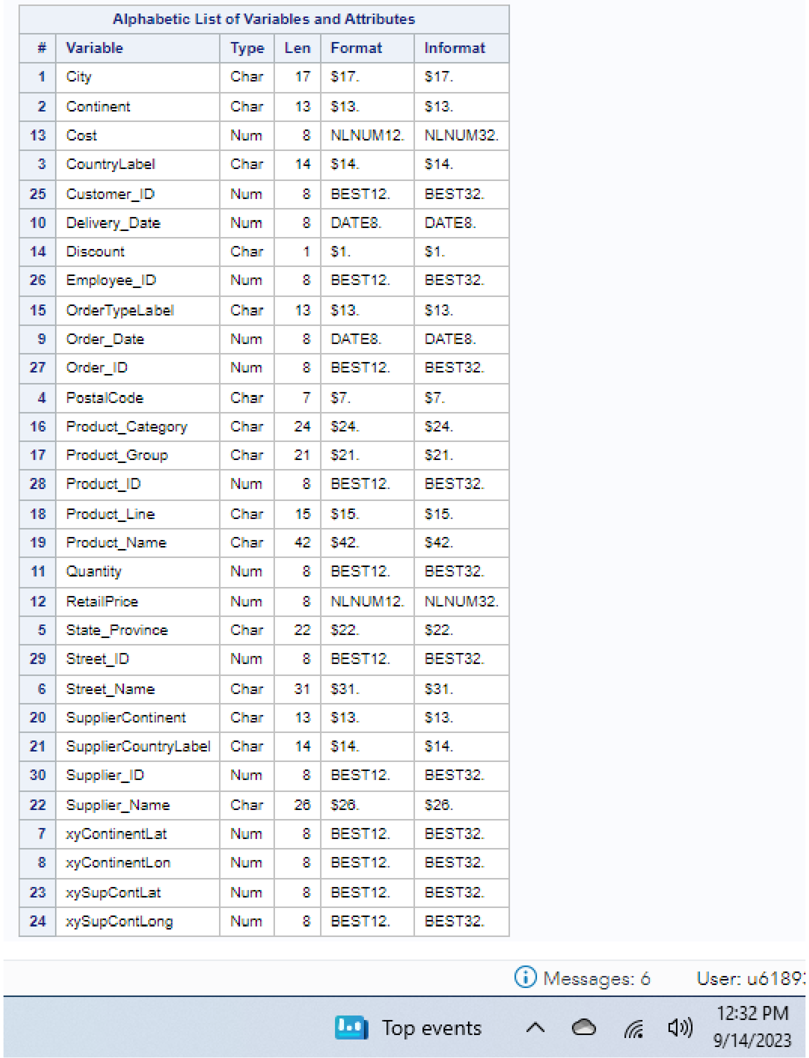
☑ Generate SAS variable names

Start reading data at row:

The CONTENTS Procedure

Data Set Name	SASUSER.IMPORT	Observations	951669
Member Type	DATA	Variables	30
Engine	V9	Indexes	0
Created	09/14/2023 12:22:05	Observation Length	400
Last Modified	09/14/2023 12:22:05	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			

Figure 2: Products dataset variable list and data type information.



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.

Now, I create a business question for what the company is trying to understand whether products are selling accurately. And study a few critical factors that help make a business very profitable. Used summary and descriptive statistics analysts as a continuous variable is “Quantity” with one categorical variable, ‘ Product Category,’ which will help to see the average order size, the most popular product category, or the variability in customer preferences.

Figure 3: The pie chart shows the volume of orders by category in SAS Studio.

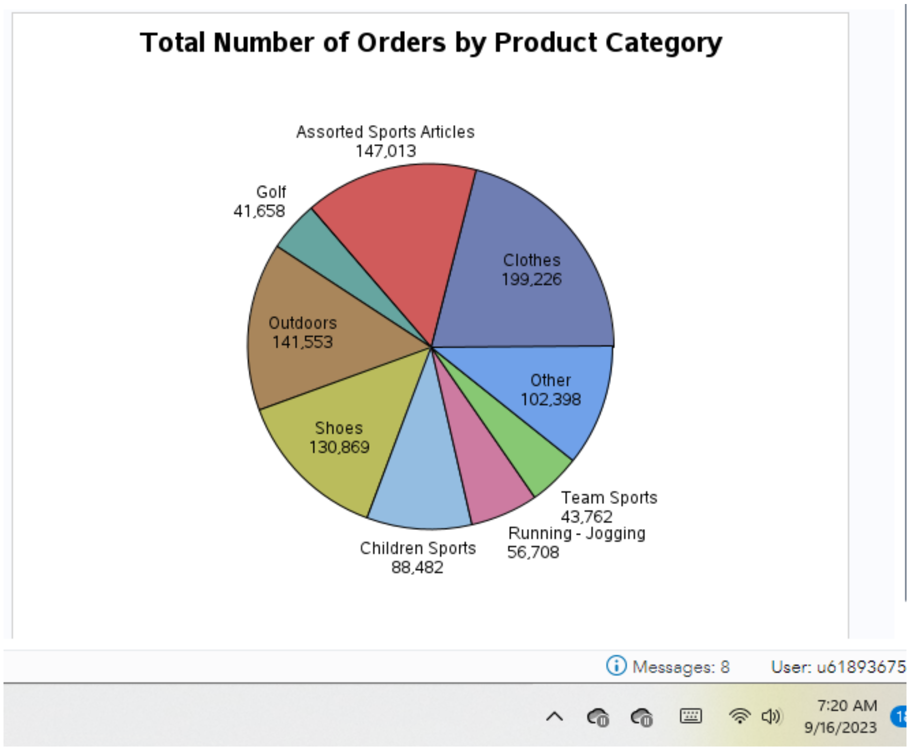


Figure 4: A bar chart to show the number of orders by customer country label in SAS Studio.

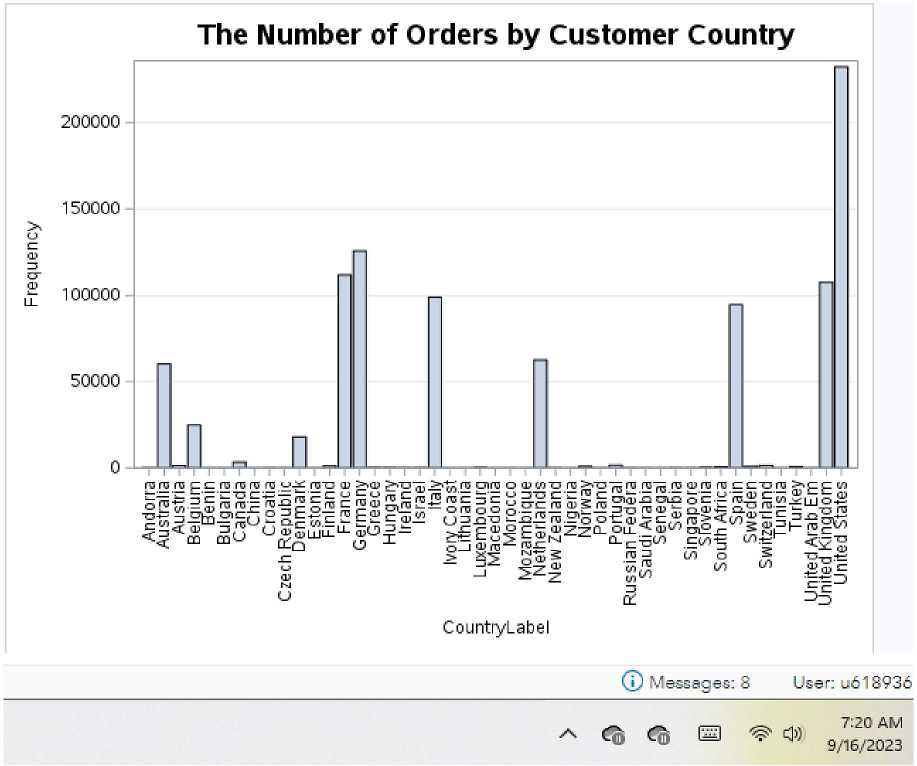


Figure 5: A histogram to show the percentage of orders by delivery time in SAS Studio.

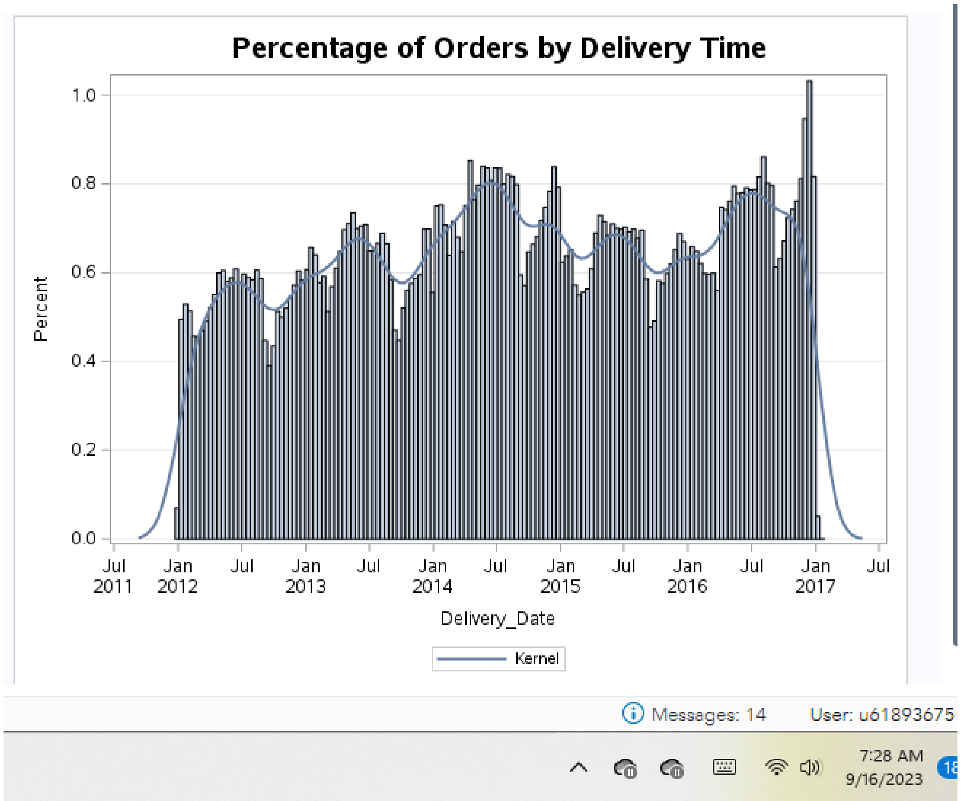
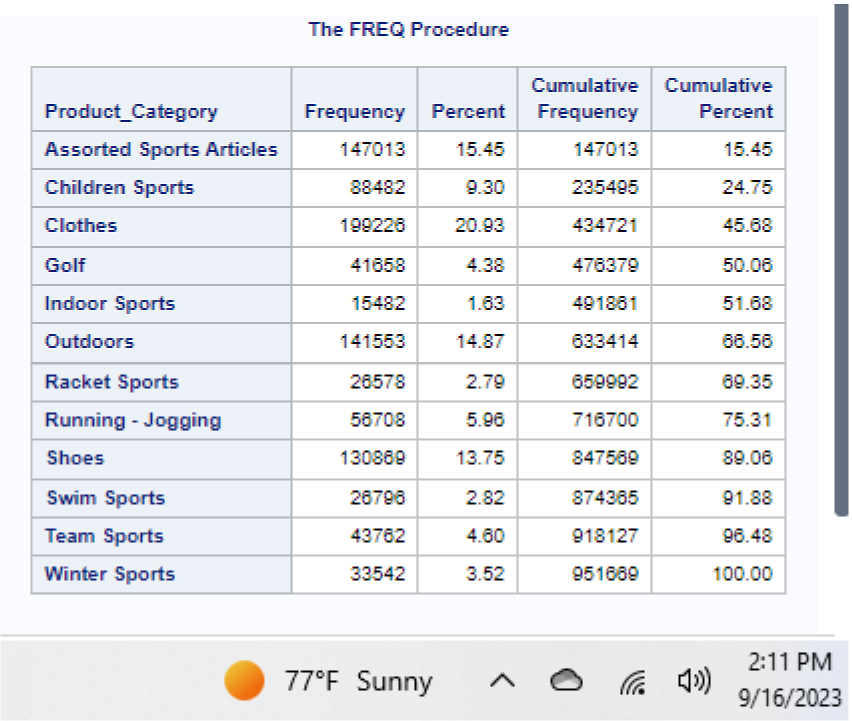


Figure 6: One-Way Frequencies analysis by 'Product_Category' in SAS Studio.



The FREQ Procedure

Product_Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Assorted Sports Articles	147013	15.45	147013	15.45
Children Sports	88482	9.30	235495	24.75
Clothes	199226	20.93	434721	45.68
Golf	41658	4.38	476379	50.06
Indoor Sports	15482	1.63	491861	51.68
Outdoors	141553	14.87	633414	66.56
Racket Sports	26578	2.79	659992	69.35
Running - Jogging	56708	5.96	716700	75.31
Shoes	130869	13.75	847569	89.06
Swim Sports	26796	2.82	874365	91.88
Team Sports	43762	4.60	918127	96.48
Winter Sports	33542	3.52	951669	100.00

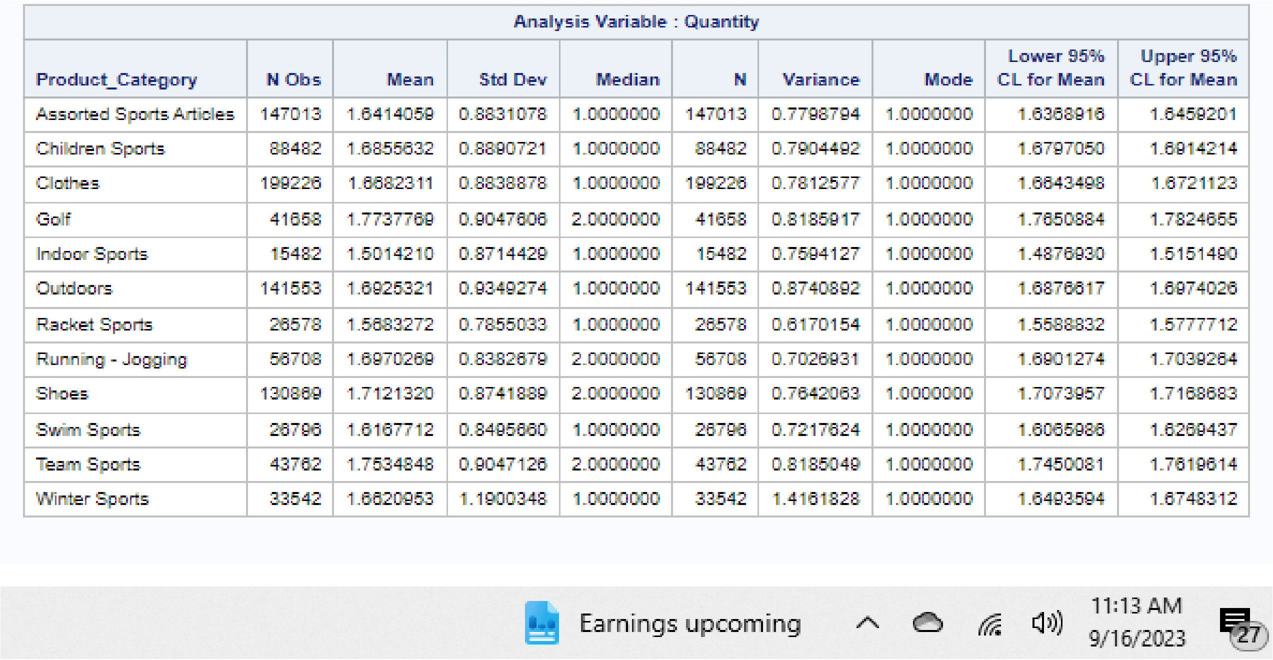
Mean

The first thing to check the mean value is to see an average data point for each variable. Outliers and high values affect the mean values. For example, in the products dataset, 'Golf' has the highest mean of 1.77, and the 'Indoor Sports' variable has the lowest average product category quantity of 1.50. Also, if the continuous variable has a few missing values, the mean value might be great to replace with the missing values.

Median

The median is preferably measured at a distance from the mean if there are outliers or high variability in the dataset. The data is usually distributed if the mean and median length are small. The median represents the middle value of a given variable list data value, so outliers and high values do not affect the median. Thus, the median of all the product category sales quantities is 1 or 2, and there is a small gap between the mean and the median value for the product category variables. Moreover, if the variable has many missing values, it may be great to use the median value.

Figure 7: Output Summary Statistics under the Statistic task and use product_Category dataset in SAS Studio.



Analysis Variable : Quantity

Product_Category	N Obs	Mean	Std Dev	Median	N	Variance	Mode	Lower 95% CL for Mean	Upper 95% CL for Mean
Assorted Sports Articles	147013	1.6414059	0.8831078	1.0000000	147013	0.7798794	1.0000000	1.6368916	1.6459201
Children Sports	88482	1.6855632	0.8890721	1.0000000	88482	0.7904492	1.0000000	1.6797050	1.6914214
Clothes	199226	1.6682311	0.8838878	1.0000000	199226	0.7812577	1.0000000	1.6643498	1.6721123
Golf	41658	1.7737769	0.9047606	2.0000000	41658	0.8185917	1.0000000	1.7650884	1.7824655
Indoor Sports	15482	1.5014210	0.8714429	1.0000000	15482	0.7594127	1.0000000	1.4876930	1.5151490
Outdoors	141553	1.6925321	0.9349274	1.0000000	141553	0.8740892	1.0000000	1.6876617	1.6974026
Racket Sports	26578	1.5683272	0.7855033	1.0000000	26578	0.6170154	1.0000000	1.5588832	1.5777712
Running - Jogging	56708	1.6970269	0.8382679	2.0000000	56708	0.7026931	1.0000000	1.6901274	1.7039264
Shoes	130869	1.7121320	0.8741889	2.0000000	130869	0.7642063	1.0000000	1.7073957	1.7168683
Swim Sports	26796	1.6167712	0.8495660	1.0000000	26796	0.7217624	1.0000000	1.6065986	1.6269437
Team Sports	43762	1.7534848	0.9047126	2.0000000	43762	0.8185049	1.0000000	1.7450081	1.7619614
Winter Sports	33542	1.6620953	1.1900348	1.0000000	33542	1.4161828	1.0000000	1.6493594	1.6748312

Mode

The mode is the most frequent data point that is rarely used, primarily as if categorical or class variables have missing values and can be replaced with mode values. The mode value of all 'Product_Category' variables is 1, which primarily shows orders of one product for each category.

Variance and Distribution

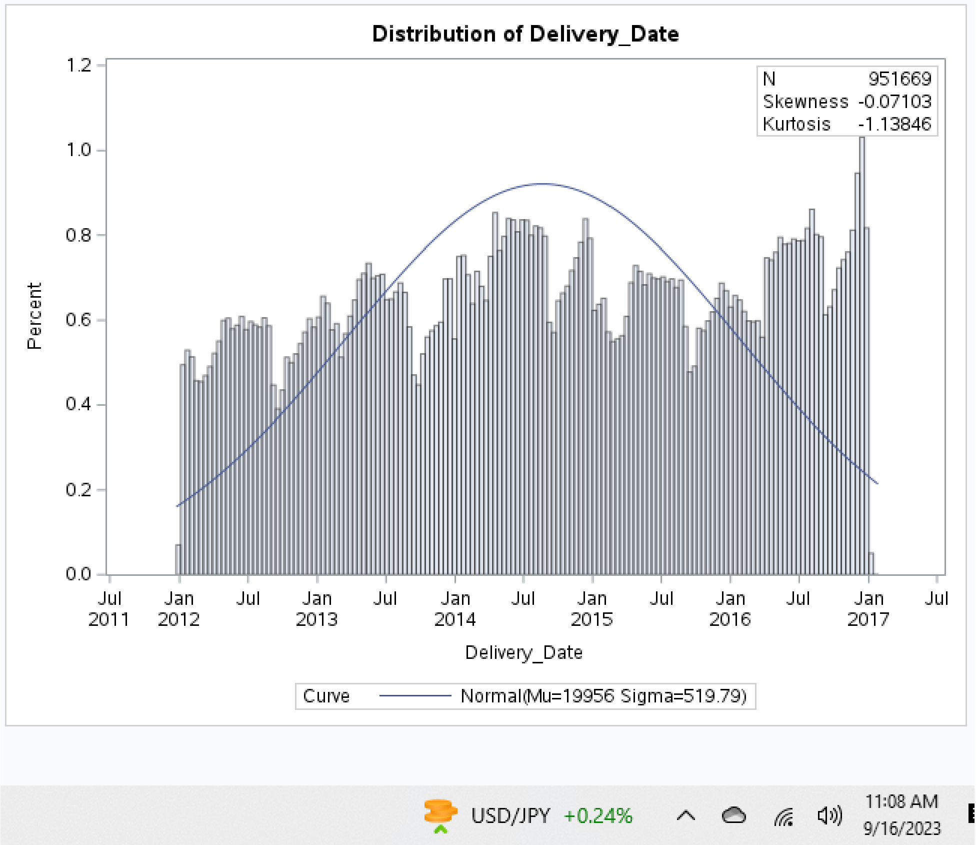
The result of variance and distribution shows how data points spread that measure range differ between the minimum and maximum values. Outliers and high values affect the content. The most helpful measure is the sample variance, the average squared deviations of each observation from the mean. Another standard deviation measure is the square root of the variance and is in the same units of measurement as the original data.

Figure 7 shows the 'Product_Category' sale quantity summary statistics. The 'Winter Sports' variance was higher than other product categories, with a standard deviation of 1.19, a data point far from the mean.

Lower and Upper 95% confidence

That helps to see how data points spread inside the 95% confidence. All the categories have a close value for upper and lower 95% confidence, so all the data points are close to each other.

Figure 8: The result of Distribution Analysis by 'Delivery_Date' in SAS Studio.



Skewness

Skewness tells that the dataset has an asymmetric distribution or is not symmetrical. There are three different distributions. One zero skew means the distribution is balanced (mean=median). Another negative skew is when the skewness value is negative (mean<median), and a positive skew is when the skewness value is positive(mean>median). For example, the skewness of the delivery order distribution analysis result shows a negative number around -0.07, which is also so close to the zero point, which is symmetrical.

Kurtosis

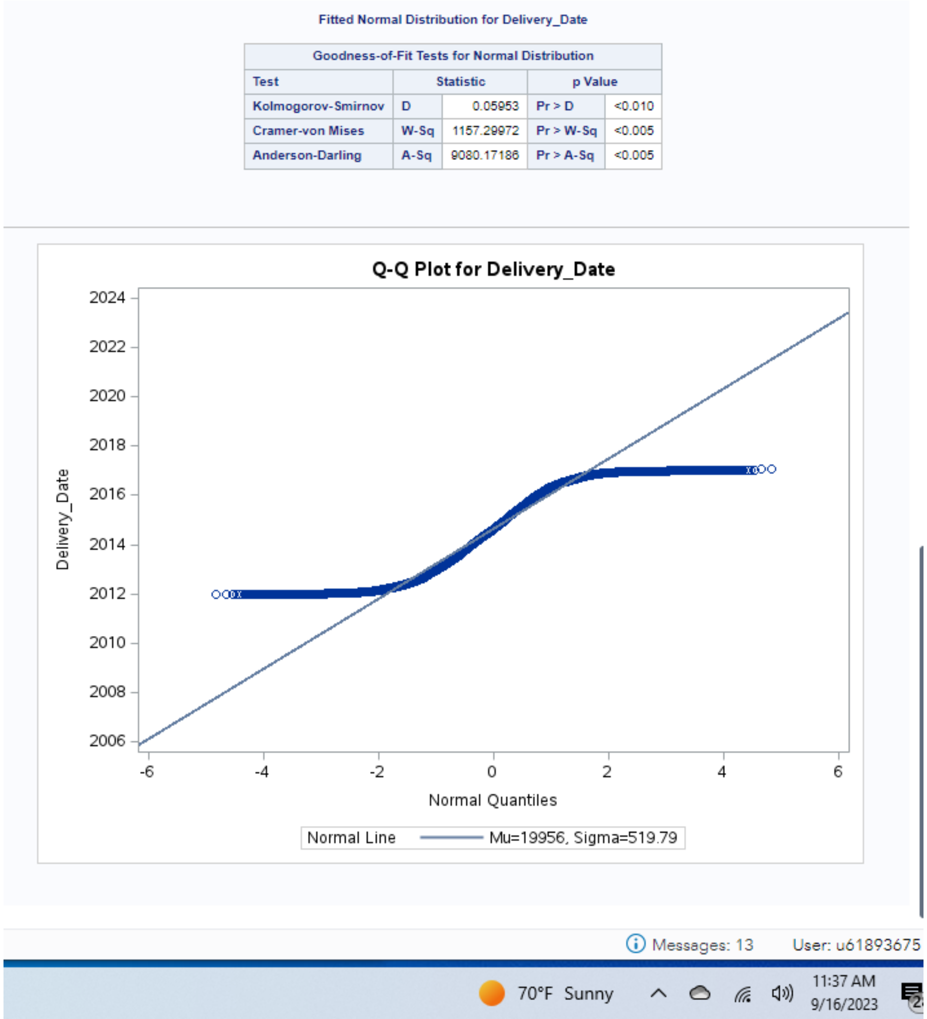
Kurtosis shows the variable's probability or frequency, which also helps to compare which variable has a heavy distribution tail with three kurtosis types. So many different ranges people use. Use zero for the normal kurtosis distribution because the best explanation for the case outliers is the number of zeros for kurtosis. Medium tails are **mesokurtic(kurtosis=0)**, low kurtosis is **platykurtic(kurtosis<0)** distribution is flatter and has smaller tails, and high kurtosis is **leptokurtic(kurtosis>0)** has a sharper peak than a bell shape**.** The order by delivery distribution analysis results show kurtosis -1.13846, which is platykurtic, so the delivery date dataset might have outliers.

Interpretation

The standard curve shows that the distribution is rightly skewed. A goodness-of-fit test is how well the sample data set fits an entire population with normal distribution and another way to say how target values are related to the independent values in a model. Kolmogorov-Smirnov test applies a large sample of over 2000. The other two tests also have the same reason to use, which helps to know whether the example of the normal distribution. Our three p-values are lower than 0.05, which means the data set is not a normal distribution.

The Q – Q plot shows how data points fall on a straight line. We discussed that the previous kurtosis result shows the dataset might have outliers, and the Q-Q chart proves it. The delivery date variable has outliers' data points close to the line, and the data does not fit straight.

Figure 9: The resulting distribution analysis is a histogram of the 'Delivery_Date' variable on SAS Studio.

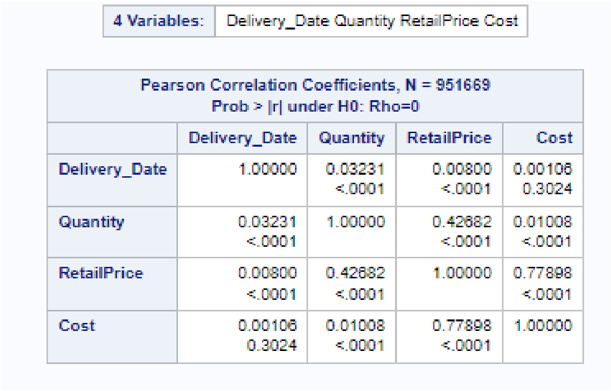


Fitted Normal Distribution for Delivery_Date

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.05953	Pr > D	<0.010
Cramer-von Mises	W-Sq	1157.29972	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	9080.17186	Pr > A-Sq	<0.005

Correlation Analysis

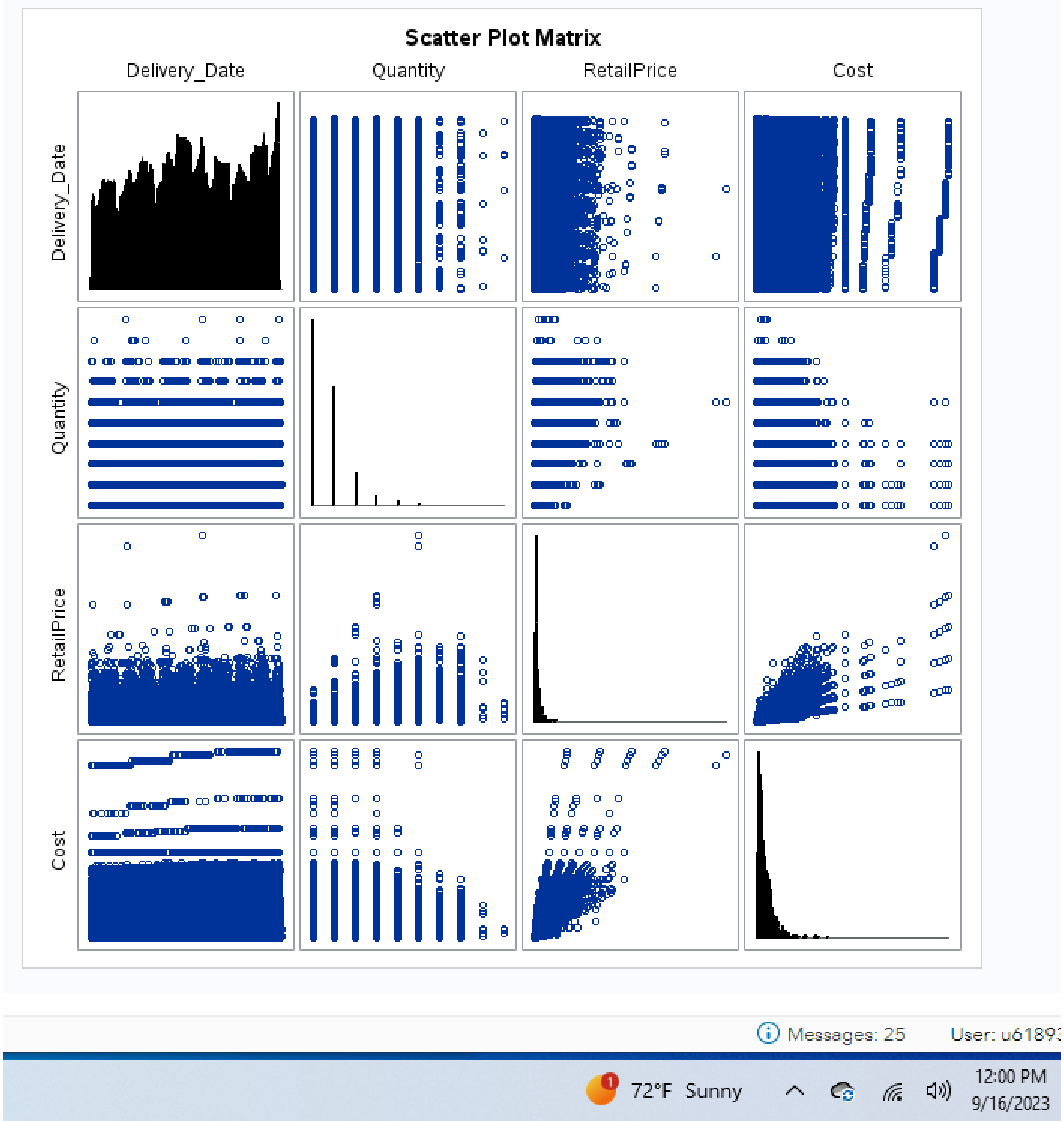
Figure 10: The result of correlation analysis in SAS Studio.



4 Variables:	Delivery_Date Quantity RetailPrice Cost

Pearson Correlation Coefficients, N = 951669
Prob > |r| under H0: Rho=0

	Delivery_Date	Quantity	RetailPrice	Cost
Delivery_Date	1.00000	0.03231 <.0001	0.00800 <.0001	0.00106 0.3024
Quantity	0.03231 <.0001	1.00000	0.42682 <.0001	0.01008 <.0001
RetailPrice	0.00800 <.0001	0.42682 <.0001	1.00000	0.77898 <.0001
Cost	0.00106 0.3024	0.01008 <.0001	0.77898 <.0001	1.00000



The first table shows the p-value positive number, and the scatter plot shows nonrelation except cost and retail price, which have a positive correlation because the first table result is 0.77, a good number, which means a 77% positive relation. Thus, if the variables have a high relation that might affect the model result, retail price and cost might be removed before the applied model.

Conclusion

I used the sample data set to create descriptive analytics with the products.csv data set, with 951669 intakes representing aggregated data and 30 attributions. I made descriptive statistics with continuous variables Quantity and Delivery_Date. The ‘Delivery_Date’ sample data set does not fit a distribution because the Q-Q Plot for the Delivery_Date chart shows the dataset is not on the standard distribution line. However, the p-value is lower than 0.05 on the goodness-of-fit test table. Thus, removing the outliers from the delivery date dataset can fit the distribution line. The highest percentage of the Product category is ‘Clothes’ (%20.81), and the lowest is ‘Indoor Sports’ (%1.46). The product category ‘Golf’s mean (1.77) is more than the product category Team Sports’s mean (1.75), as we discussed above in the mean explanation; however, we can’t say product category ‘Golf’ sells better than the product category ‘Team Sports’ because the outlier affects product category Team Sports’ mean. Max portion sales for the product category ‘Clothes’ is 199,226, and for the category’ Assorted Sports Articles,’ it is 147,013. Both product categories, ‘Clothes’ and ‘Assorted Sports Articles,’ have a positive skew (mean>median).

References

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

ISBN: 9781954844476.

Kishore, A. (2023). Descriptive Analytics: Steps, Techniques, Use Case, Examples.

https://www.knowledgehut.com/blog/data-science/descriptive-analytics

Cote, C. (2021). What Is Descriptive Analytics? 5 Examples.

https://online.hbs.edu/blog/post/descriptive-analytics

Bhandari, P. (2023). Descriptive Statistics/ Definitions, Types, Examples.

https://www.scribbr.com/statistics/descriptive-statistics/

Sharma, S. (2019). Descriptive Statistics. Horizons University.

https://www.researchgate.net/publication/333220406_Descriptive_Statistics