Customer Data Analysis Report

Didem B. Aykurt

Colorado State University Global

MIS543: Enterprise Performance Management

Dr. John Marlowe

October 22, 2023

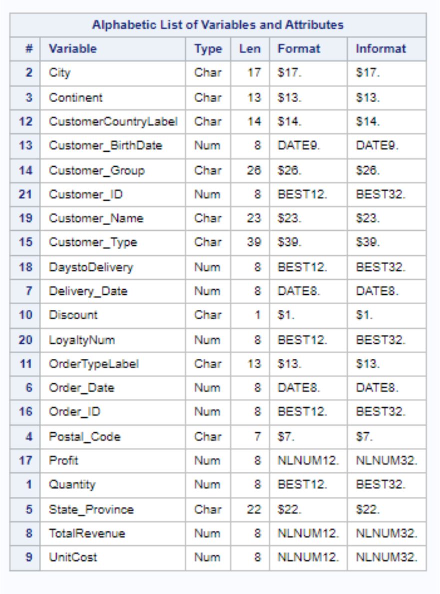
A Data Analysis of Sales and Customers

This paper contains hypothesis testing and descriptive and predictive analysis of sales and customers. It uses data to understand a company’s current and potential customers, needs, preferences, behavior, competitors, products, and market opportunities. It can help segment their customers, identify pain points, tailor their products and services, spot trends and prospects, differentiate the company’s business from competitors, and optimize marketing efforts. The steps involved in a data analysis of sales and customers are:

Perform market research to find customers: The company needs to gather information about the size, revenue, standards, and external factors of its industry. Also, they need to identify their ideal customers, demographics, psychographics, and buying behavior.

Perform competitive analysis to find a market advantage: The company needs to understand who their direct and indirect competitors are, what their strengths and weaknesses are, how they position themselves in the market, and what their unique selling propositions are.

The customers.csv dataset with retail, internet, and catalog sales information in five years processing data to answer company needs. First, clean the data by renaming a few variables in Excel the ‘Loyalty Num’ to ‘LoyaltyNum,’ ‘Total Revenue’ to ‘TotalRevenue,’ and Unit Cost’ to ‘UnitCost’ and import the customers.csv file into SAS Studio.



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
2	City	Char	17	$17.	$17.
3	Continent	Char	13	$13.	$13.
12	CustomerCountryLabel	Char	14	$14.	$14.
13	Customer_BirthDate	Num	8	DATE9.	DATE9.
14	Customer_Group	Char	26	$26.	$26.
21	Customer_ID	Num	8	BEST12.	BEST32.
19	Customer_Name	Char	23	$23.	$23.
15	Customer_Type	Char	39	$39.	$39.
18	DaystoDelivery	Num	8	BEST12.	BEST32.
7	Delivery_Date	Num	8	DATE8.	DATE8.
10	Discount	Char	1	$1.	$1.
20	LoyaltyNum	Num	8	BEST12.	BEST32.
11	OrderTypeLabel	Char	13	$13.	$13.
6	Order_Date	Num	8	DATE8.	DATE8.
16	Order_ID	Num	8	BEST12.	BEST32.
4	Postal_Code	Char	7	$7.	$7.
17	Profit	Num	8	NLNUM12.	NLNUM32.
1	Quantity	Num	8	BEST12.	BEST32.
5	State_Province	Char	22	$22.	$22.
8	TotalRevenue	Num	8	NLNUM12.	NLNUM32.
9	UnitCost	Num	8	NLNUM12.	NLNUM32.

Business Question and Hypothesis

Question 1:

Which sales channel (catalog, online, or retail) has the highest sales amount?

Null Hypothesis (H0): There is no significant difference in sales among the three channels.

Alternative Hypothesis (H1): There is a significant difference in sales amount among the three channels.

Question 2:

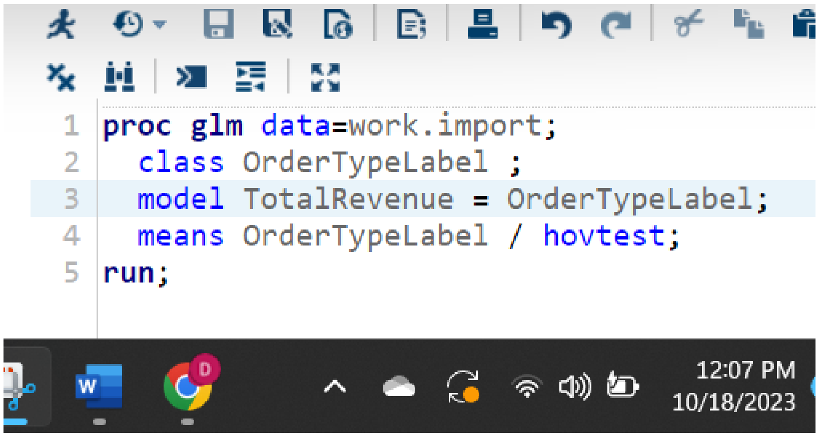
What is the relationship between the customer groups and sales amount?

Null Hypothesis (H0): There is no significant relationship between the customer groups and sales amount.

Alternative Hypothesis (H1): There is an essential relationship between customer groups and sales amount.

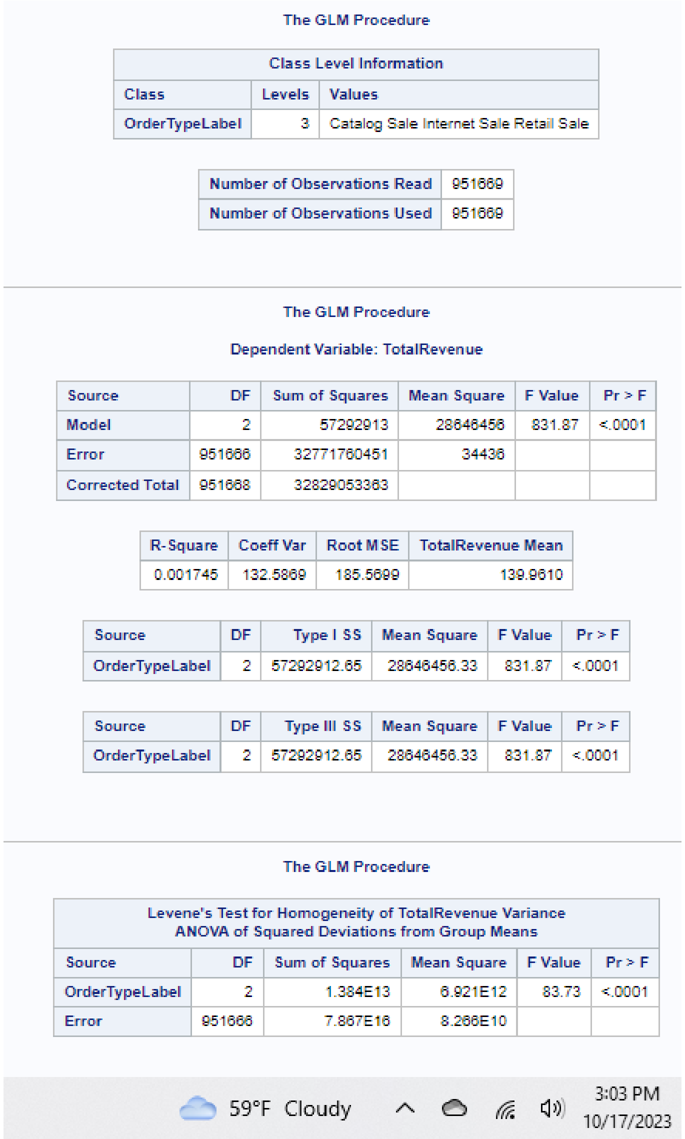
Perform a Statistical test ANOVA.

Question 1: compare the sales amount of three channels: catalog, online, and retail. To do this, we need to use a one-way ANOVA test, a generalization of the t-test for multiple groups. A one-way ANOVA test will tell if there is a significant difference in the sales amount among the three sales channels overall.



```
proc glm data=work.import;
  class OrderTypeLabel ;
  model TotalRevenue = OrderTypeLabel;
  means OrderTypeLabel / hovtest;
run;
```

Open the Tasks and Utilities pane and select Poer and Sample Size> One-Way ANOVA. This will open the One-Way ANOVA task window. Choose the data set from the Data tab. In the Options tab, choose various options for your ANOVA test, such as significance level, confidence interval, post-hoc tests, plots, and output statistics. Also, the PROC GLM function helps to create an ANOVA test.



The GLM Procedure

Class Level Information		
Class	Levels	Values
OrderTypeLabel	3	Catalog Sale Internet Sale Retail Sale

Number of Observations Read	951669
Number of Observations Used	951669

The GLM Procedure

Dependent Variable: TotalRevenue

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	57292913	28646456	831.87	<.0001
Error	951666	32771760451	34436		
Corrected Total	951668	32829053363			

R-Square	Coeff Var	Root MSE	TotalRevenue Mean
0.001745	132.5869	185.5699	139.9610

Source	DF	Type I SS	Mean Square	F Value	Pr > F
OrderTypeLabel	2	57292912.65	28646456.33	831.87	<.0001

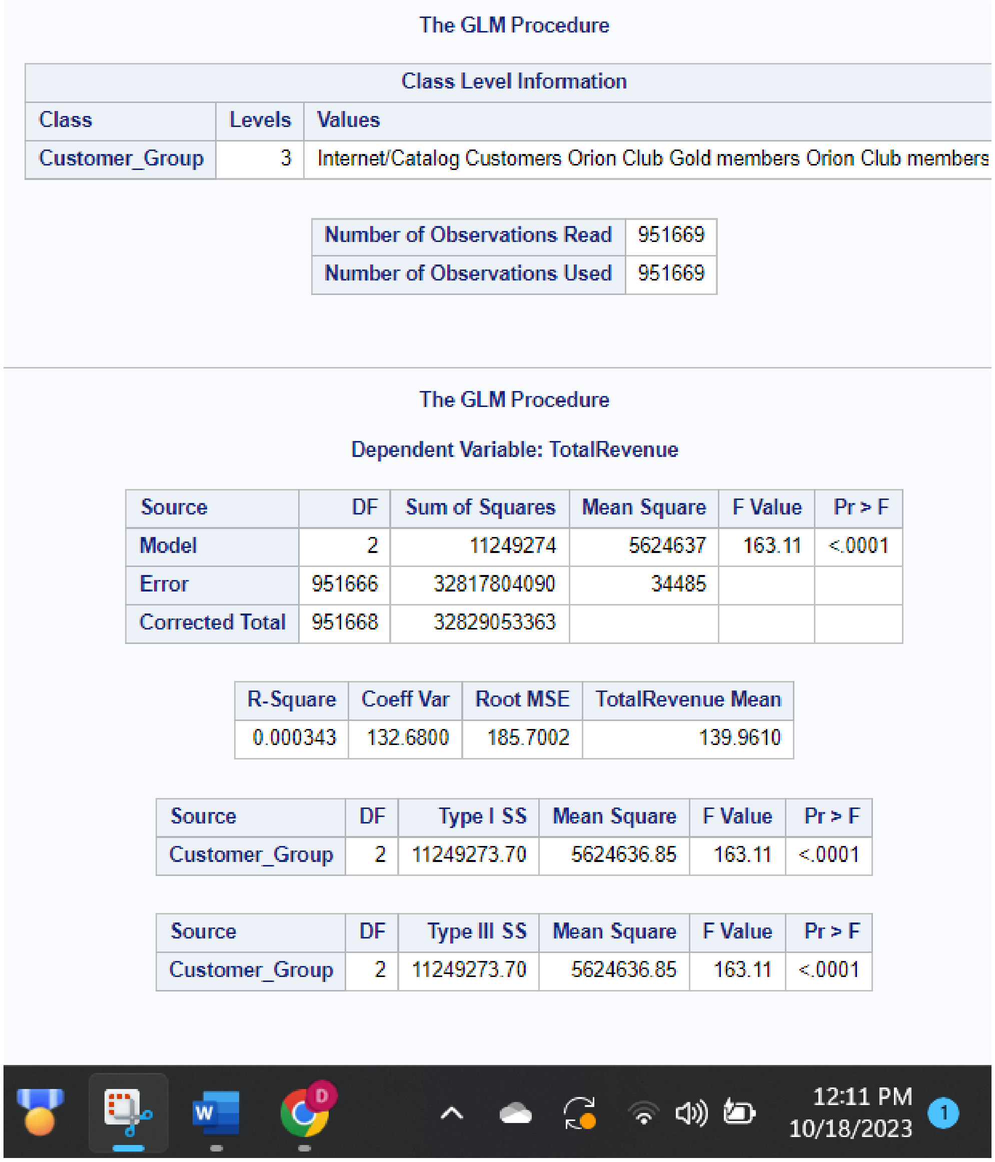
Source	DF	Type III SS	Mean Square	F Value	Pr > F
OrderTypeLabel	2	57292912.65	28646456.33	831.87	<.0001

The GLM Procedure

Levene's Test for Homogeneity of TotalRevenue Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
OrderTypeLabel	2	1.384E13	6.921E12	83.73	<.0001
Error	951666	7.867E16	8.266E10		

The ANOVA results suggest that the Order type catalog, internet, and retail sales significantly differ in sales amount among the three channels, as the (Pr>F) at 0.0001 is lower than the significance level of 0.05. This means that there is strong evidence to reject the null hypothesis. The R-squared value of .0017 indicates that only a tiny portion of the variability in "OrderTypeLabel" is explained by the "TotalRevenue."

Question 2: Compare the three different groups of customers with total revenue. ANOVA stands for Analysis of Variance. It is a statistical method to analyze differences among group means in a sample. ANOVA tests the hypothesis that the help of two or more populations is equal, generalizing the t-test to more than two groups as the customer group dataset has three groups: Internet/Catalog Customers, Orion Club Gold members, and Orion Club members.



The GLM Procedure

Class Level Information		
Class	Levels	Values
Customer_Group	3	Internet/Catalog Customers Orion Club Gold members Orion Club members

Number of Observations Read	951669
Number of Observations Used	951669

The GLM Procedure

Dependent Variable: TotalRevenue

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11249274	5624637	163.11	<.0001
Error	951666	32817804090	34485		
Corrected Total	951668	32829053363			

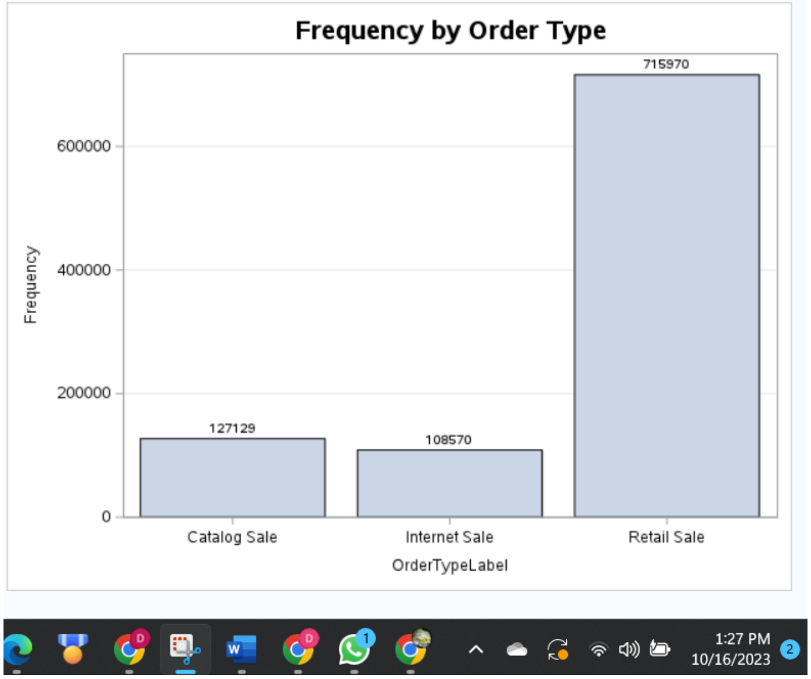
R-Square	Coeff Var	Root MSE	TotalRevenue Mean
0.000343	132.6800	185.7002	139.9610

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Customer_Group	2	11249273.70	5624636.85	163.11	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Customer_Group	2	11249273.70	5624636.85	163.11	<.0001

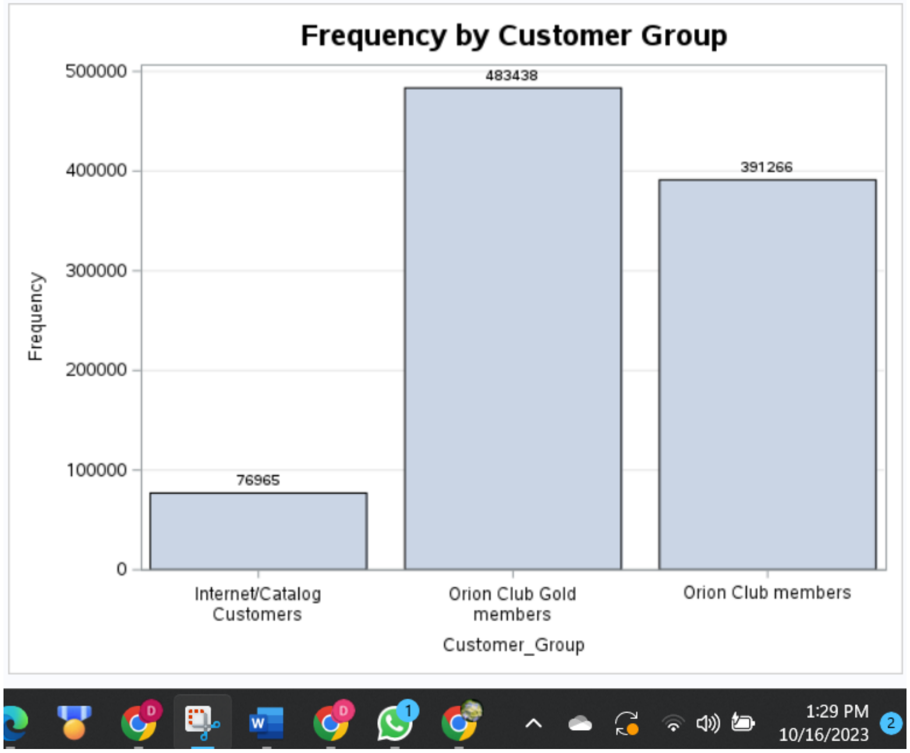
The ANOVA result suggests that the ‘Customer_Group’ variable has a significant relationship between customer groups and sales amount as the p-value (Pr>F) at .0001 is lower than the significance level of .005. The R-squared value indicates that only a tiny portion of the variability in 'TotalRevenue' is explained by the 'Customer_Group' variable.

Descriptive Statistics



The "Frequency by Order Type" shows the frequency of three different types of orders: "Catalog Sale," "Internet Sale," and "Retail Sale". The y-axis represents the frequency, ranging from 0 to 600,000, while the x-axis represents the order type. The "Retail Sale" bar is the tallest, with a frequency of 715,970. The "Internet Sale" bar is the second tallest, with a frequency of 108,570. The "Internet Sale" bar is the second tallest, with a frequency of 108,570. The bar for "Catalog Sale" is the shortest, with a frequency of 127,129.

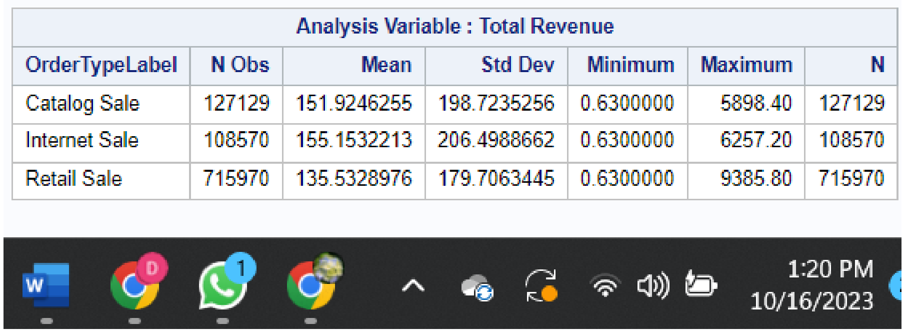
This graph helps compare the frequencies of different types of sales. It shows retail sales are much more frequent than internet and catalog sales. Internet sales are more frequent than catalog sales.



The “Frequency by Customer Group” shows the frequency of three different customer groups: “Internet/Catalog Customers,” “Orion Club Gold members,” and “Orion Club members.” The y-axis represents the frequency, while the x-axis represents the customer group. The bar for “Orion Club Gold members” is the tallest, with a frequency of 483,438. The bar for “Orion Club members” is in the middle, with a frequency of 391,266. The bar for “Internet/Catalog Customers” is the shortest, with a frequency of 76,965.

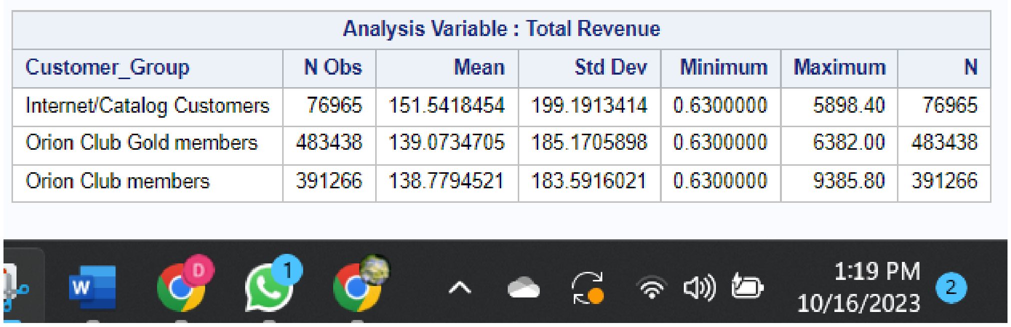
This graph helps compare the frequencies of different customer groups. It shows that Orion Club Gold members are much more frequent than Orion Club members and Internet/Catalog Customers.

Figure 1: Result of statistic summary total revenue by order type in SAS Studio.



Analysis Variable : Total Revenue						
OrderTypeLabel	N Obs	Mean	Std Dev	Minimum	Maximum	N
Catalog Sale	127129	151.9246255	198.7235256	0.6300000	5898.40	127129
Internet Sale	108570	155.1532213	206.4988662	0.6300000	6257.20	108570
Retail Sale	715970	135.5328976	179.7063445	0.6300000	9385.80	715970

The statistical results compare the three order types. The mean value can give us an idea of the average value of each order type, such as catalog sales' average revenue at $151.92, Internet sales at $155.15, and retail sales at $135.53. The standard deviation can give us an idea of the variation in each order type. The highest standard deviation is Internet sales at 206.499, Catalog sales at 198.72, and Retail sales at 179.71. The minimum and maximum values can give us an idea of the range of values for each order type; three category sales have the same minimum value, and Retail sales have the highest value of maximum number at $9585.80.

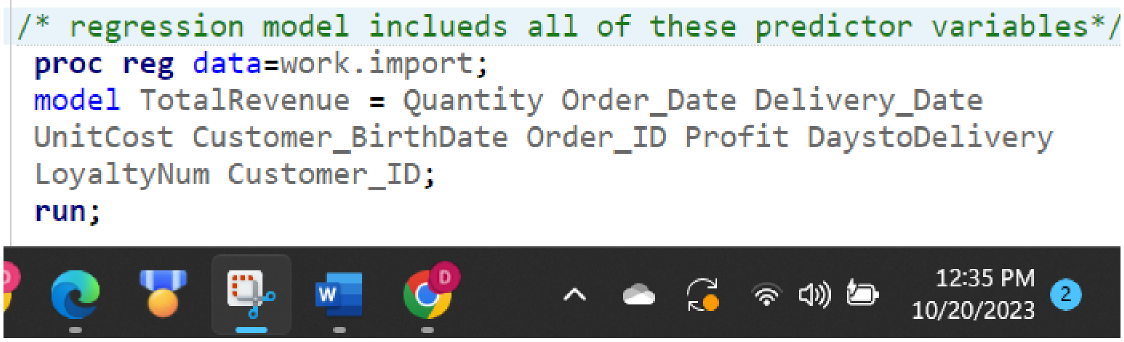


Analysis Variable : Total Revenue						
Customer_Group	N Obs	Mean	Std Dev	Minimum	Maximum	N
Internet/Catalog Customers	76965	151.5418454	199.1913414	0.6300000	5898.40	76965
Orion Club Gold members	483438	139.0734705	185.1705898	0.6300000	6382.00	483438
Orion Club members	391266	138.7794521	183.5916021	0.6300000	9385.80	391266

The table shows that Internet/Catalog Customers have the highest average total revenue, with a mean of $151.541. Orion Club members have the lowest average real income, with a standard of $138.779. Orion Club Gold members have the lower average total revenue, with a mean of $139.0734.

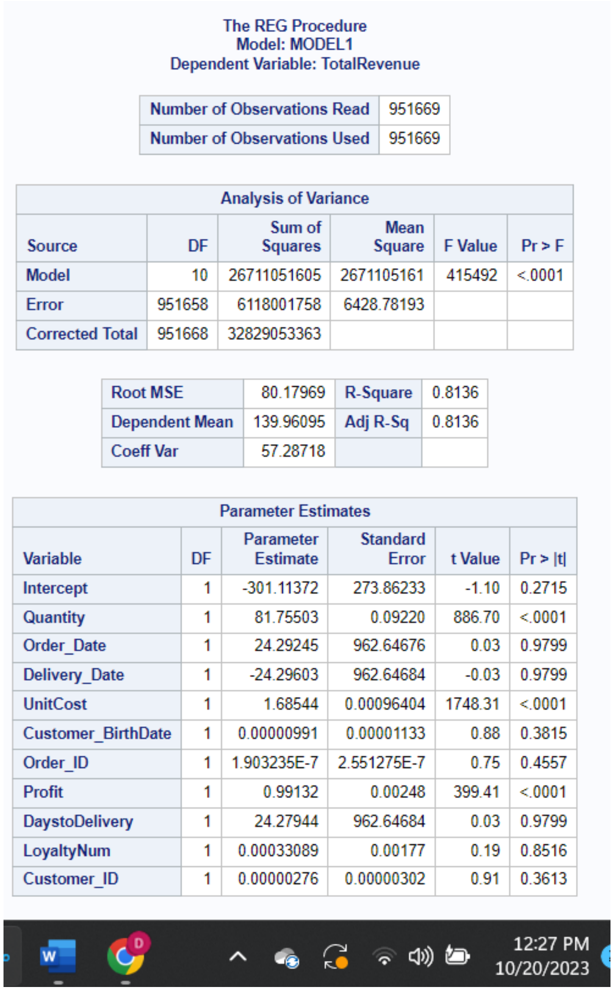
Predictive Analysis

The aim is to find which predictor variables can increase ‘TotalRevenue.’ Predictor variables are ‘Quantity,’ ‘Order_Date,’ ‘Delivery_Date,’’ UnitCost,’ ‘Customer_Birthdate, ‘Order_ID,’ ‘Profit,’ ‘DaystoDelivery,’ ‘LoyaltyNum,’ and ‘Customer_ID.’ The Proc Reg function helps to create a model to predict.



```
/* regression model includeds all of these predictor variables*/
 proc reg data=work.import;
 model TotalRevenue = Quantity Order_Date Delivery_Date
 UnitCost Customer_BirthDate Order_ID Profit DaystoDelivery
 LoyaltyNum Customer_ID;
 run;
```

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables. It is used to determine how much the value of the dependent variable changes when any of the independent variables are varied. The Analysis of Variance table contains metrics p-value and R-square that explain how the dataset fits the model as Pr>F shows .0001, which is lower than the significant level of 0.05; it indicates that the results are not due to change and is likely to be true. R-square is a measure of how closely the predicted values match the actual values in the data set. The most common measure of dataset fit is the coefficient of determination (R-squared). This statistical measure represents the proportion of the variance in the dependent variable explained by the model's independent variable(s) as a table result shows an R-squared value of 0.81, indicating that all of the variance in the dependent variable is explained by the independent variable(s) in the model.



The REG Procedure
Model: MODEL1
Dependent Variable: TotalRevenue

Number of Observations Read	951669
Number of Observations Used	951669

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	26711051605	2671105161	415492	<.0001
Error	951658	6118001758	6428.78193		
Corrected Total	951668	32829053363			

Root MSE	80.17969	R-Square	0.8136
Dependent Mean	139.96095	Adj R-Sq	0.8136
Coeff Var	57.28718		

Parameter Estimates

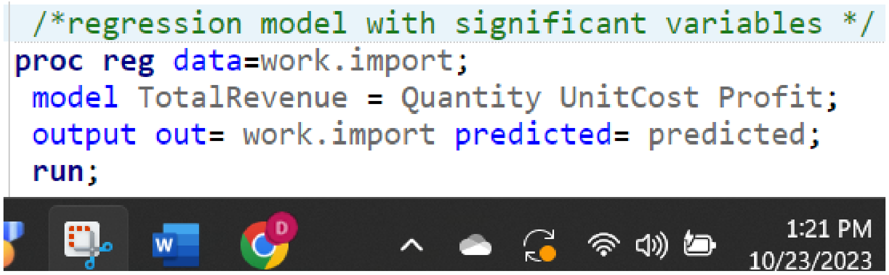
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-301.11372	273.86233	-1.10	0.2715
Quantity	1	81.75503	0.09220	886.70	<.0001
Order_Date	1	24.29245	962.64676	0.03	0.9799
Delivery_Date	1	-24.29603	962.64684	-0.03	0.9799
UnitCost	1	1.68544	0.00096404	1748.31	<.0001
Customer_BirthDate	1	0.00000991	0.00001133	0.88	0.3815
Order_ID	1	1.903235E-7	2.551275E-7	0.75	0.4557
Profit	1	0.99132	0.00248	399.41	<.0001
DaystoDelivery	1	24.27944	962.64684	0.03	0.9799
LoyaltyNum	1	0.00033089	0.00177	0.19	0.8516
Customer_ID	1	0.00000276	0.00000302	0.91	0.3613

A Parameter Estimates table is related to a regression analysis used to predict future sales based on predictor variables. The Intercept row represents the constant term in the regression equation, while the Variable rows represent the predictor variables. The Parameter Estimates column represents the estimated coefficients for each predictor variable, so the positive correlation with ‘Quantity’ of 82.75, ‘UnitCost’ of 1.68, and ‘Profit’ of 0.99. The Standard Error column represents the standard error of each coefficient estimate as those three variables (Quantity, UnitCost, Profit) have a small number of standard errors. The Pr > |t| column represents the p-value for each coefficient estimate, so significant p-value variables are ‘Quantity,’ ‘UnitCost,’ and ‘Profit’ at 0.0001.

Thus, the result of regression analysis indicates that three variables, ‘Quantity,’ ‘UnitCost,’ and ‘Profit,’ are statistically significant predictors of ‘TotalRevenue.’ However, ‘Order_Date,’ ‘Delivery_Date,’ ‘ Customer_BirthDate,’ ‘Order_ID,’ ‘DaystoDelivery,’ ‘LoyaltyNum,’ and ‘Customer_ID’ do not appear to be significant predictors in this model.

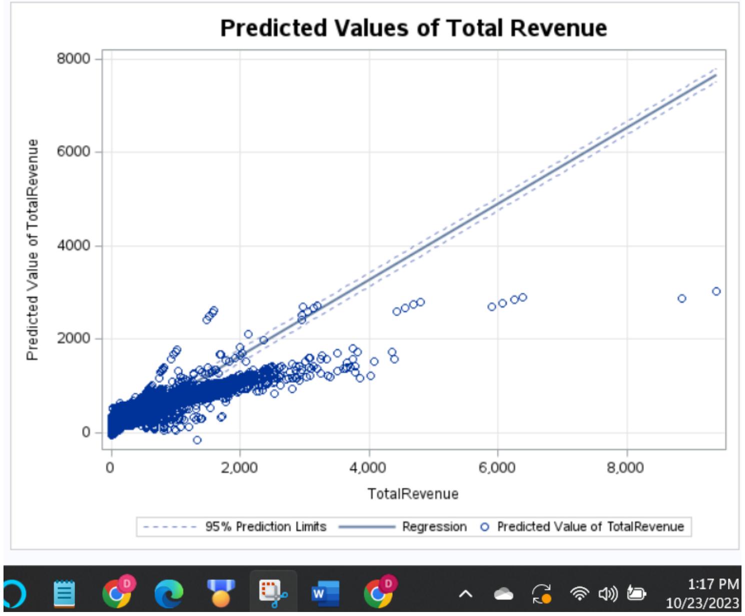
Applying the new linear regression analysis without non-significant variables may improve the regression result and help to see the relation between predicted value and total revenue.

Create a new predict column with a regression model on the table, then run with expected value and total revenue by scatter plot.



```
 /*regression model with significant variables */
proc reg data=work.import;
 model TotalRevenue = Quantity UnitCost Profit;
 output out= work.import predicted= predicted;
 run;
```

The scatterplot with a regression line represents the predicted values from the multiple linear regression model. This visualization and analysis aim to assess the performance of the regression model to compare the total sales value with the predicted values. The customer type of orders is mostly retail rather than catalog and internet orders. Customer group as



Conclusion

By analyzing historical customer data and other relevant data points, hypothesis testing and predictive analytics help identify patterns and trends that can help them make informed decisions about company operations. Most order types are Retail orders. The company focuses on the retail department, improving customer service and supporting high-volume customers in retail charges. Another strategy is for the company to use promotions and discounts on online orders and catalog orders to increase both sale types. The highest number of Customer group type Orion Club Gold members with a frequency of 483,438 in this group of customers might increase volume that would be too organized or prepared to support this customer group. The strong relationship with sales amount ‘Quantity,’ ‘UnitCost,’ and ‘Profit’ statistically significant predictors of ‘TotalRevenue.’ All three variables have a strong positive relation with sales amount. As a scatter plot has a regression line showing the best fit on the line, the predicted value is lower than the registration line, and the prediction value is lower than the required value point.

References

Cody, R.(2021*). A Gently Introduction to Statistics Using SAS Studio in the Cloud.* SAS Institute. SAS Institute.
SAS.com (n.d.). *Simple Liner Regression.* https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect003.htm

SAS.com (n.d.). *N-Way ANOVA Task.* https://support.sas.com/documentation/cdl/en/webeditorug/68254/HTML/default/viewer.htm#n14k2j5mchdnwon1r1txjyq87hkb.htm

SAS.com (n.d.). *Summary Statistics Task.* https://support.sas.com/documentation/cdl/en/webeditorug/68254/HTML/default/viewer.htm#n1nzvhc62eedvhn17j5v1rx9y0nt.htm

Helper.com (n.d.). Customer Data Analysis- How to Analysis Data in 7 Steps. https://www.callcentrehelper.com/customer-data-analysis-steps-162911.htm

Kelley, K. (2023). What is Data Analysis?: Process, Types, Methods, and Techniques. https://www.simplilearn.com/data-analysis-methods-process-types-article