

**RETAIL SALES PERFORMANCE**

**A report analyzing sales performance retail store customer’s transactional data**

**based on product category, year, province,city and to explore which source could be the best for retail stores increase their sales and location for new business startup.**

INTRODUCTION

This report covers Retail store customer’s transactional data that were recorded from Apr,2007 to May,2008, to analyze various different products on Different dates, with 8 variables about the data in one Data Set and to utilize some of the variables from the second Data Set with Electronic products with the category EC90, having 11 variables.

OBJECTIVES:

* To understand any scope for a particular business in any province.
* To understand any particular product can be sold at a particular duration in the year .
* To understand which products has highest sales through which source

RESEARCH QUESTIONS / HYPOTHESIS:

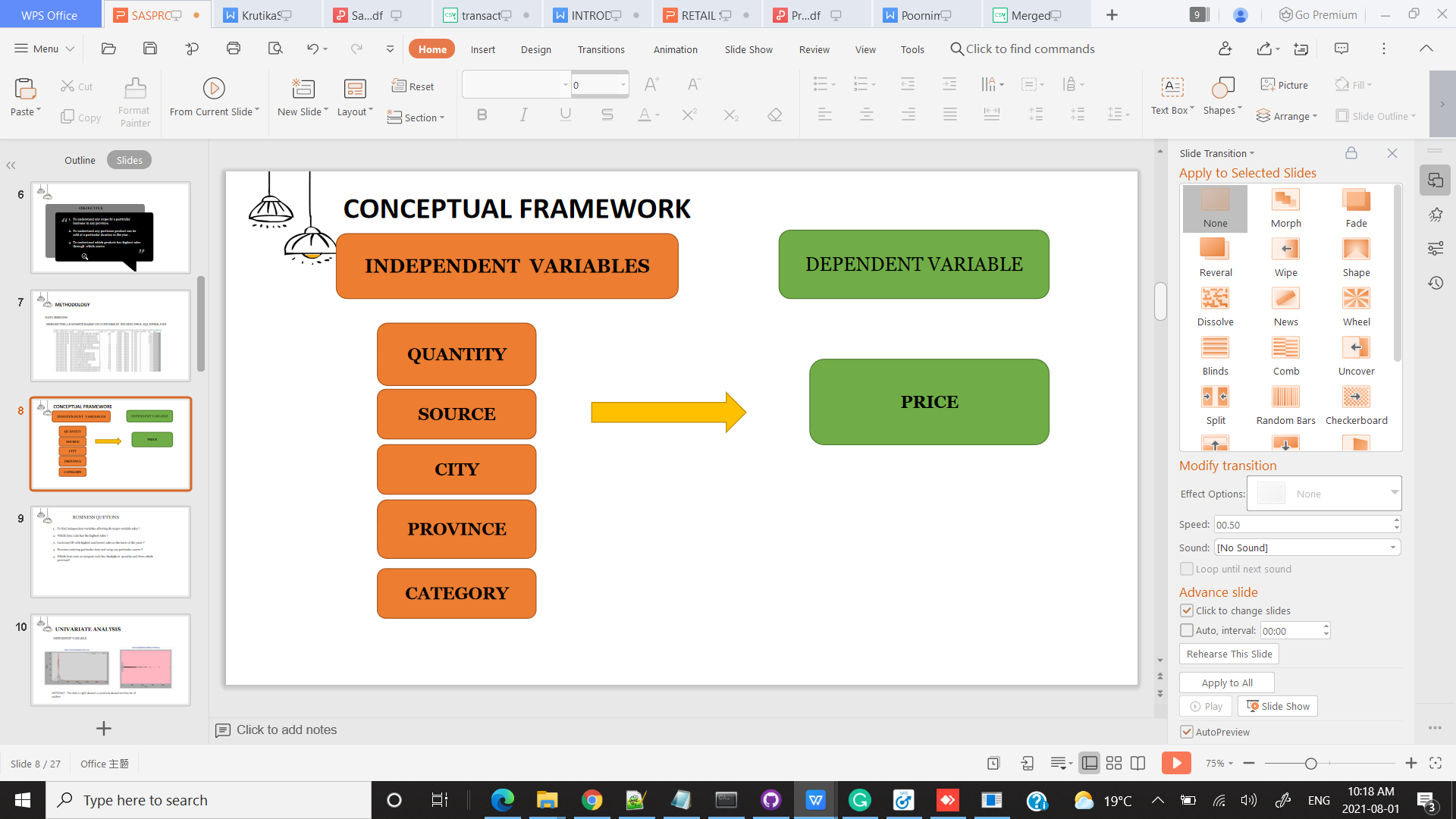
The following research question and hypothesis were used in order to build the analysis:

1. What was the effect of the Variables on sales ?

HYPOTHESIS

* Is there any association between Sales and Quantity
* Is there any association between Sales and Product Category
* Is there was an association between Sales and Province
* Is there any association between Sales and City
* Is there any association between Order Month, Source and Sum of Sales Is there any association between Order Year, Source and Sum of Sales
* Is there any association between Order Day, Source and Sum of Sales

4.CONCEPTUAL FRAME WORK



To see the effect of these independent variables on Sales.

5.METHODOLOGY

There are two datasets EC90 and customer transaction history. The Transaction data is for the year 2007-2008, operated online ,retail and through other sources of a Retail Shopping. The Items brought are in varied range products from Beds to Computers. All products are categorized and the

Quantities for the order are collected .

The other dataset EC90 from the same RETAIL Shop, gives insight about the customers purchasing only ACER ASPIRE 16" MULTIMEDIA NOTEBOOK with additional details of the orders placed from which province/city and whether the order was for first time

The variables of these dataset are as follows:

|  |  |
| --- | --- |
| EC90 | CUSTOMERTRANSACTION |
|  |  |
| OrderNumber | CustomerID |
| CustomerID | ItemCode |
| City | Source |
| Prov | OrderDate |
| PostalCode | ItemDescription |
| OrderFirstTime | Category |
| Source | Price |
| Price | Quantity |
| ItemCode |  |
| ItemDescription |  |
| Category |  |
| Quantity |  |
|  |  |
|  |  |

The variables highlighted are common in both the datasets.

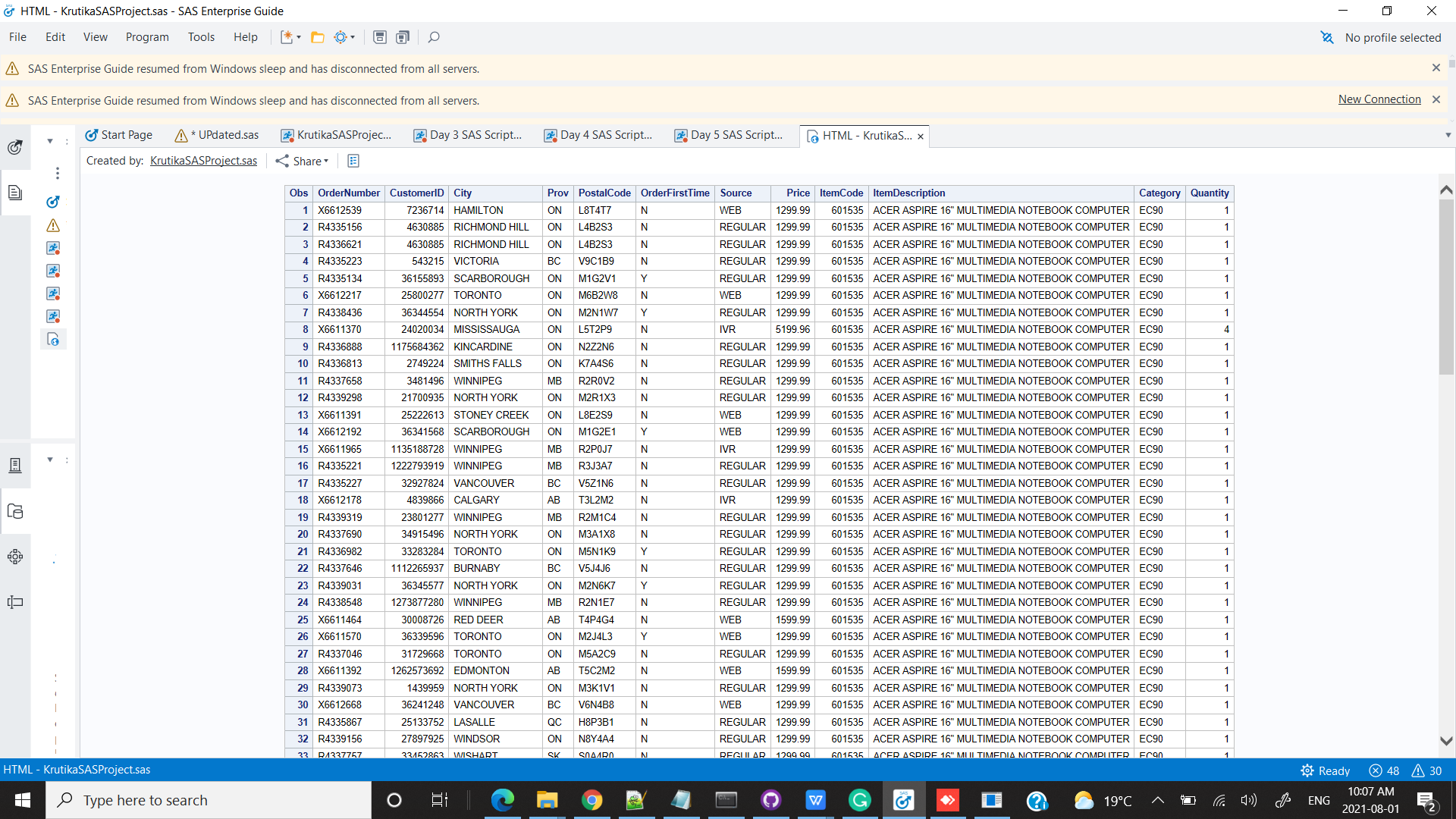
6.DATA PREPARATION

We have two datasets with some common variables, EC90 has variables like location which is not available in the other dataset.

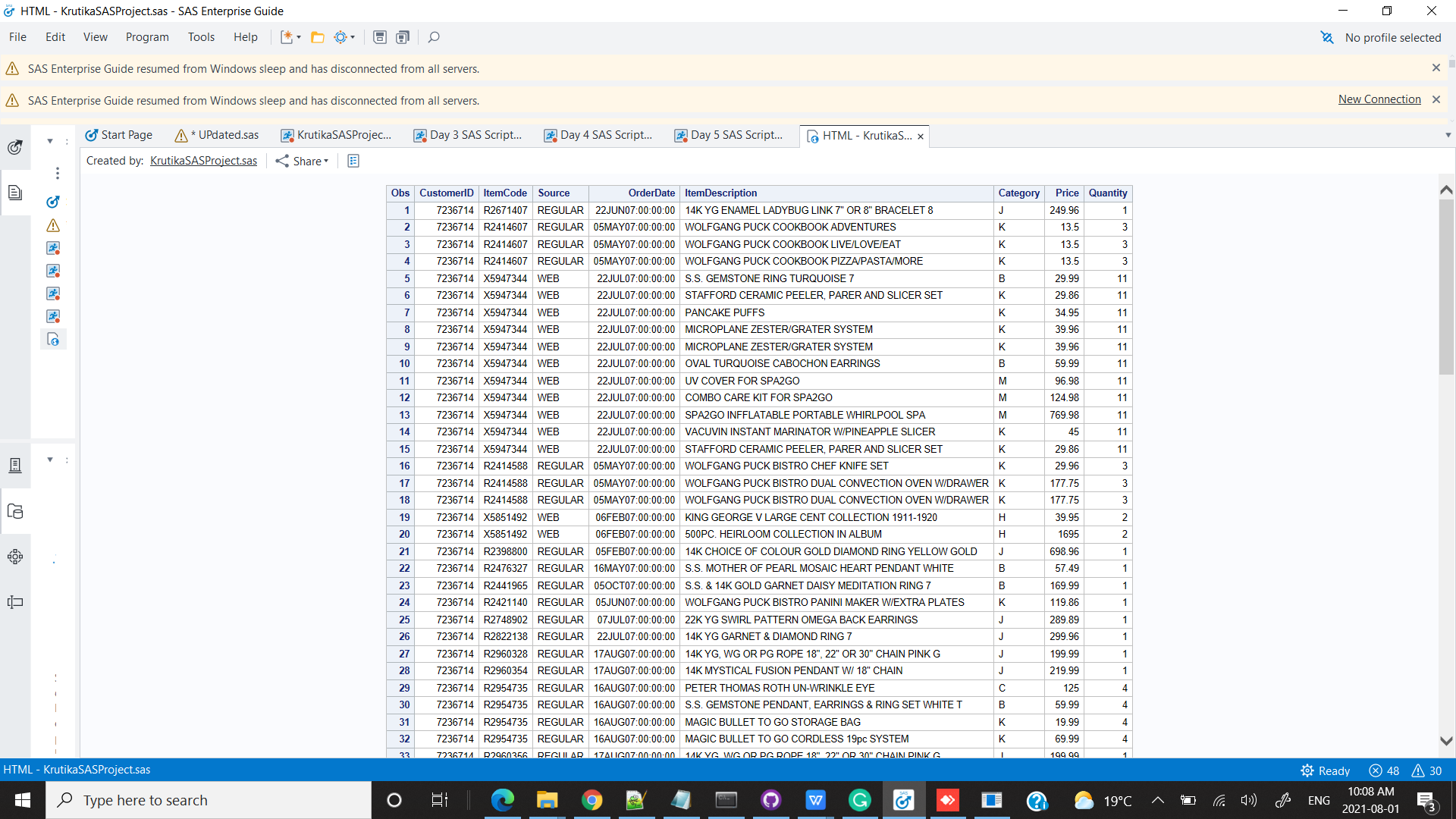
So in order to merge the dataset I did Normalization by creating a Master table with unique CustomerID and location from EC90 dataset and then I have did union of the two datasets based on Master table. So now I have two datasets with all the common variables & Orderdate from transaction history.

So after the union I get total observations as 35586 from EC90 and customer transaction data.

Ec90 DATA



CUSTOMER TRANSACTION HISTORY



/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*JOIN EC90 & TRANSACTION HISTORY WITH MASTER TABLE\*/

**proc** **sql**;

create table ksp.mergedec90transactionhistory as

select CustomerID,Price,Category,Prov,City,Quantity,Source

from ksp.ec90 as a

union all

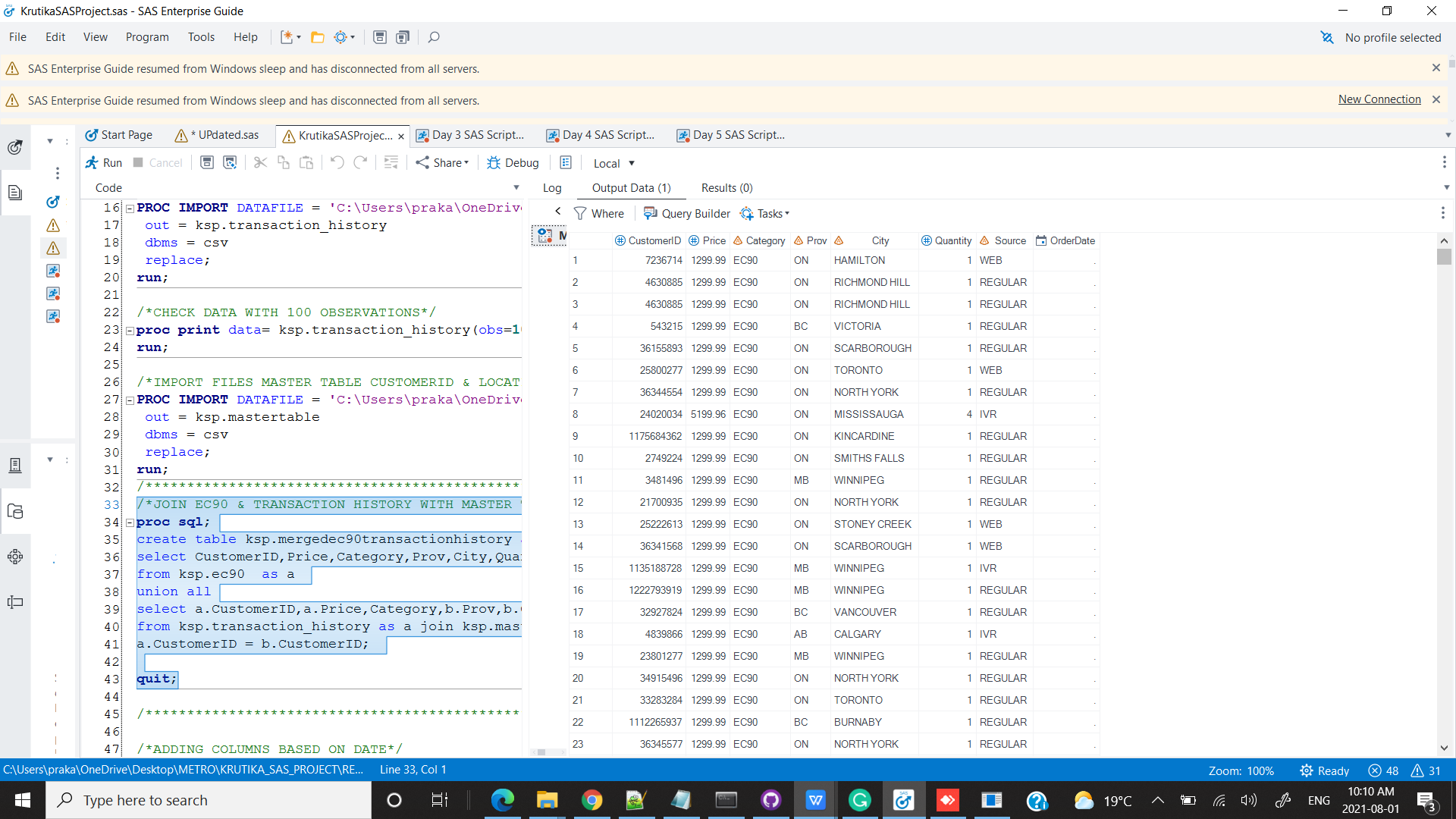
select a.CustomerID,a.Price,Category,b.Prov,b.City,a.Quantity,Source,a.OrderDate

from ksp.transaction\_history as a join ksp.mastertable as b on

a.CustomerID = b.CustomerID;

**quit**;

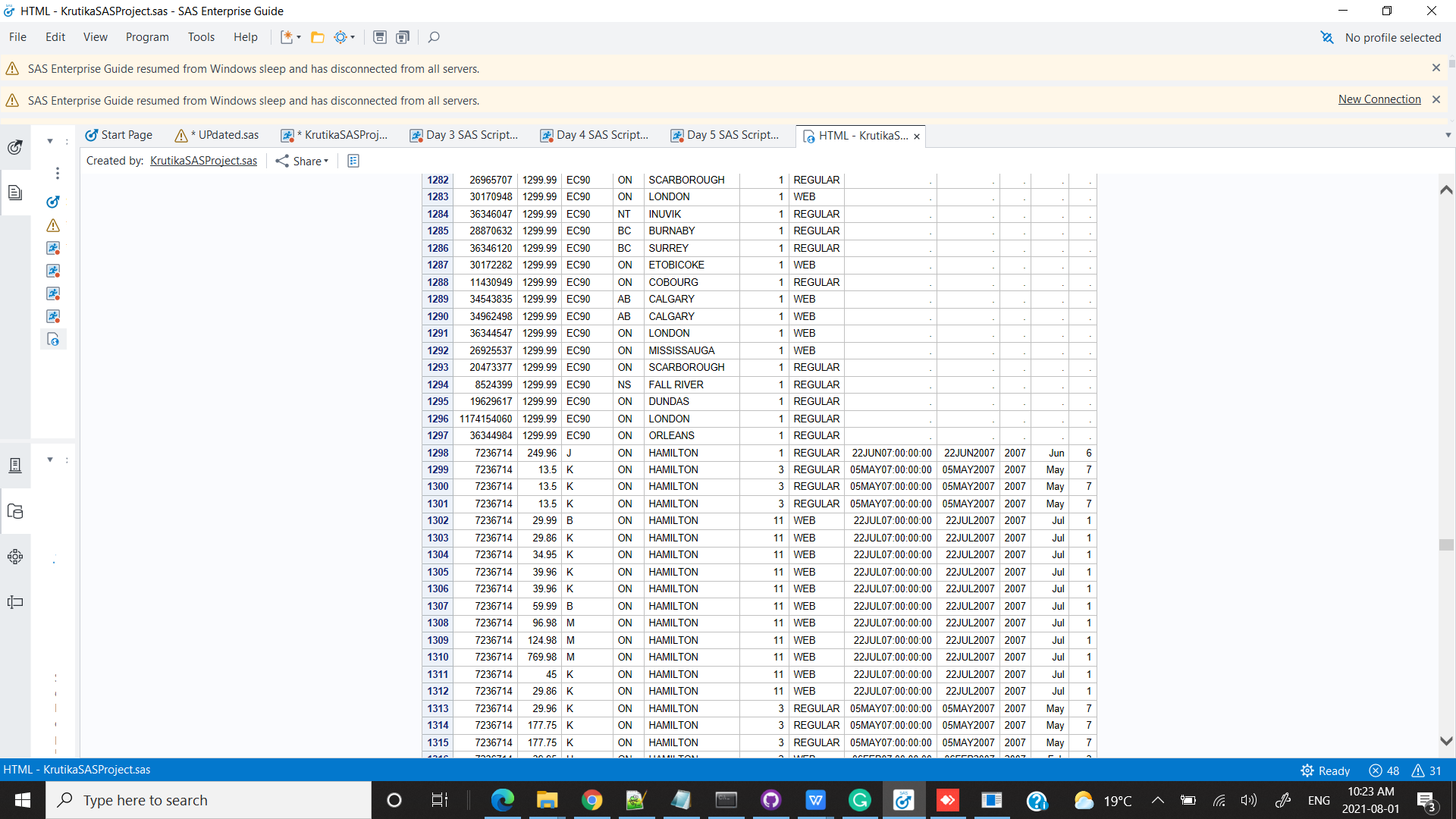
COMBINED DATA



MODIFYING ORDERDATE COLUMN

I have extracted year, month & day from the date to check the sales as per year , month and day.

NOTE : As Ec90 data has no OrderDate column it has null values.

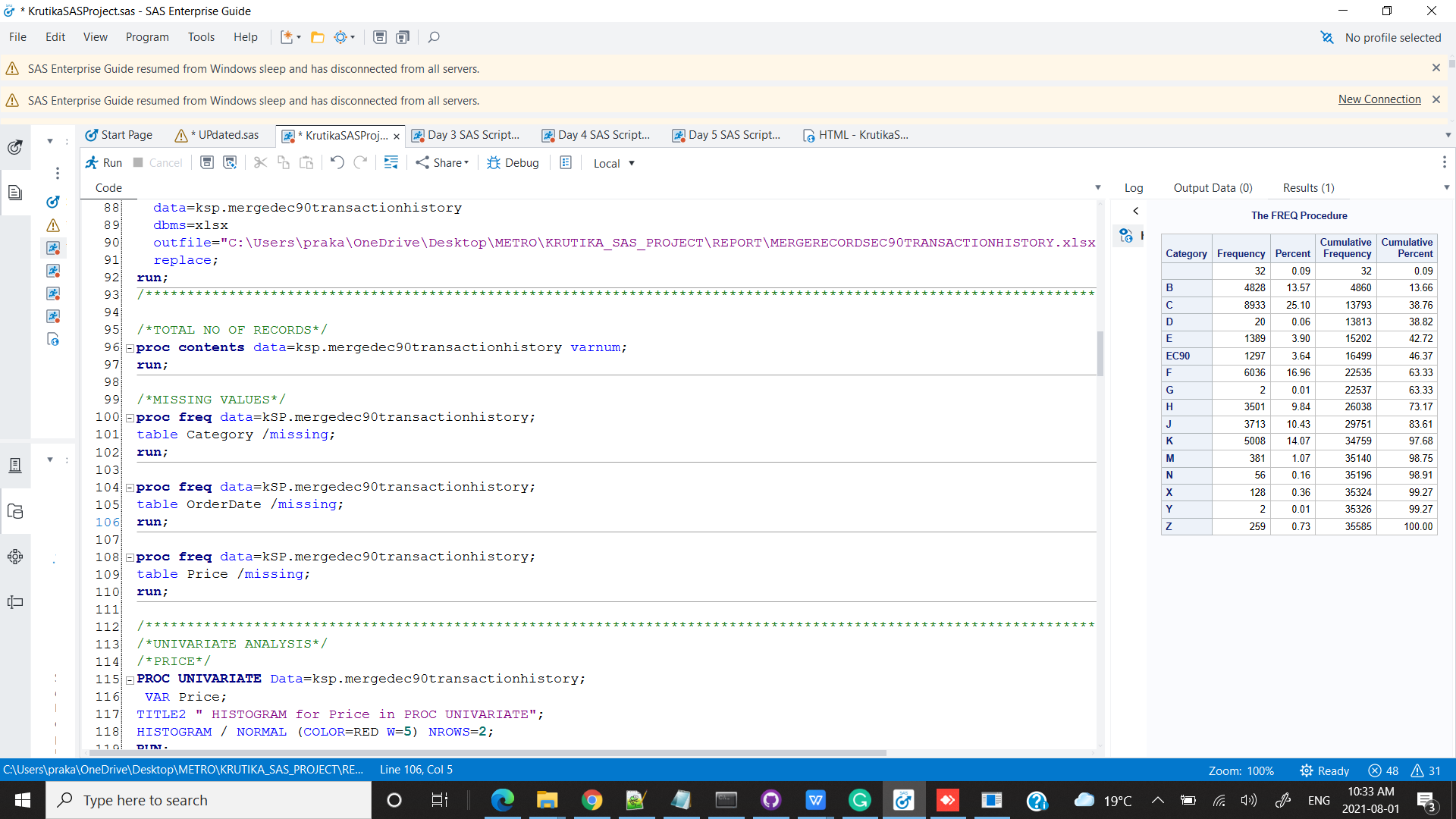


7.MISSING VALUES

**proc** **freq** data=kSP.mergedec90transactionhistory;

table Category /missing;

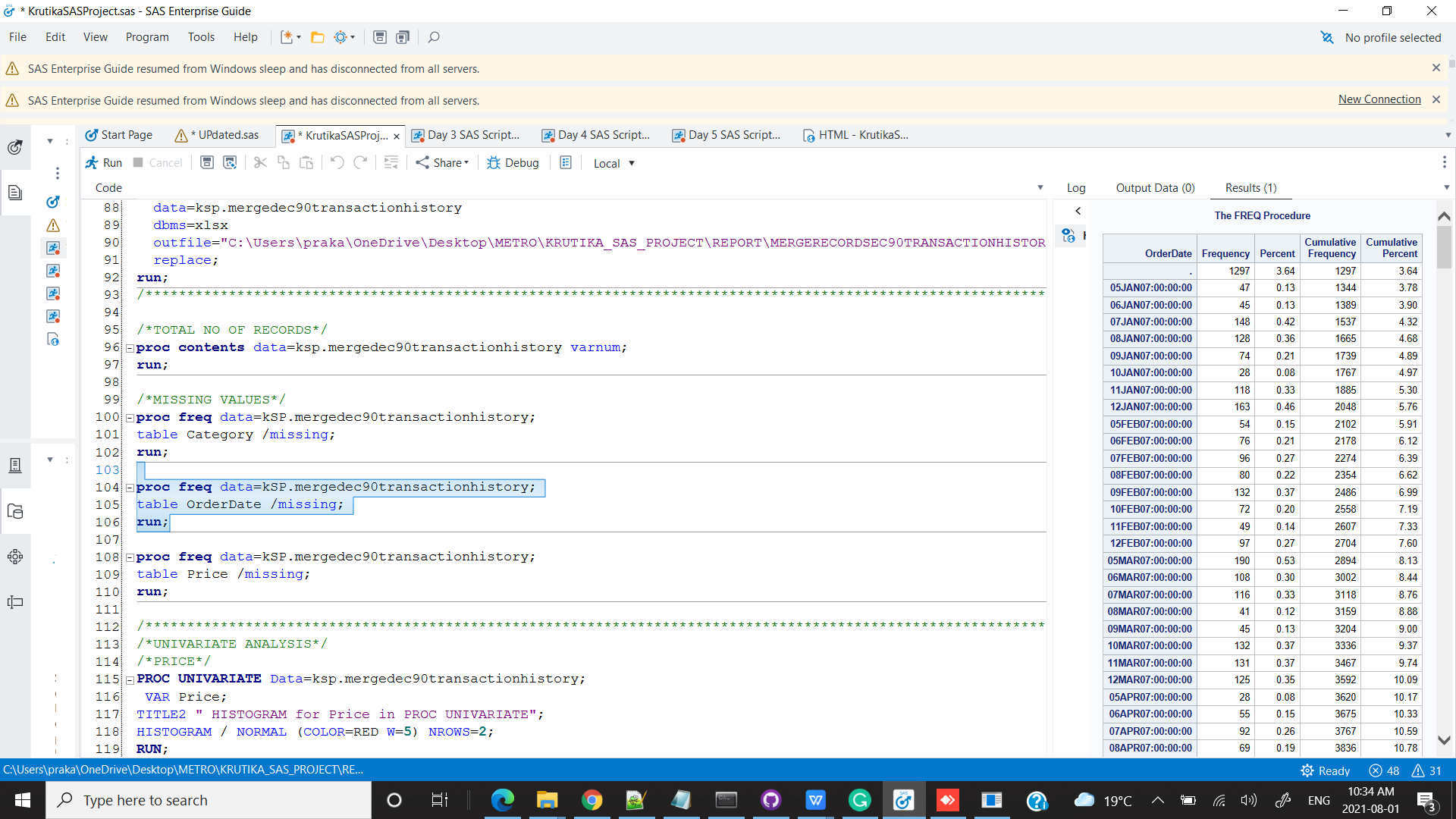
**run**;



**proc** **freq** data=kSP.mergedec90transactionhistory;

table OrderDate /missing;

**run**;



NOTE : As there is no orderDate in EC90 so all those values are null

UNIVARIATE ANALYSIS

/\*PRICE\*/

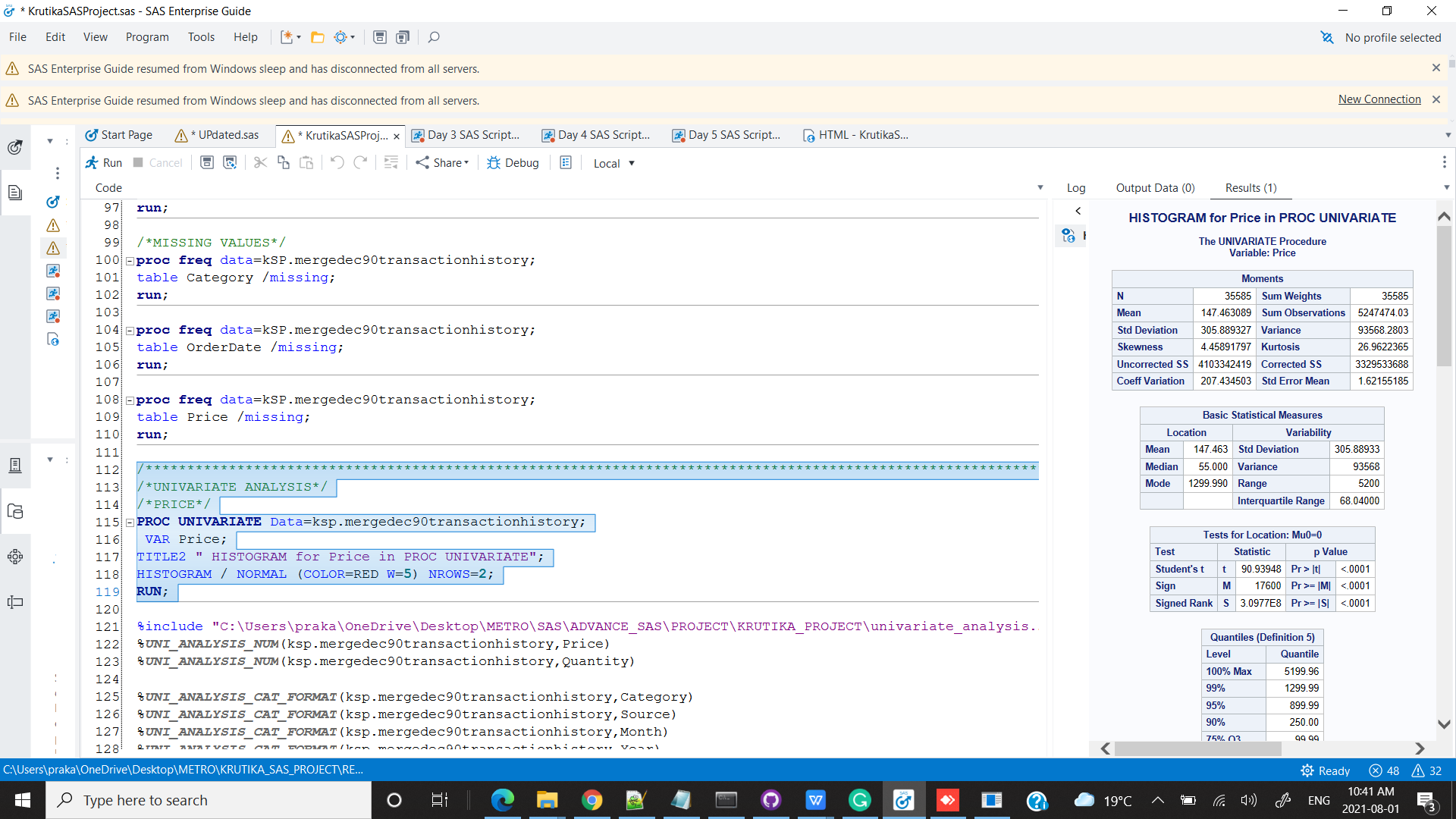
**PROC** **UNIVARIATE** Data=ksp.mergedec90transactionhistory;

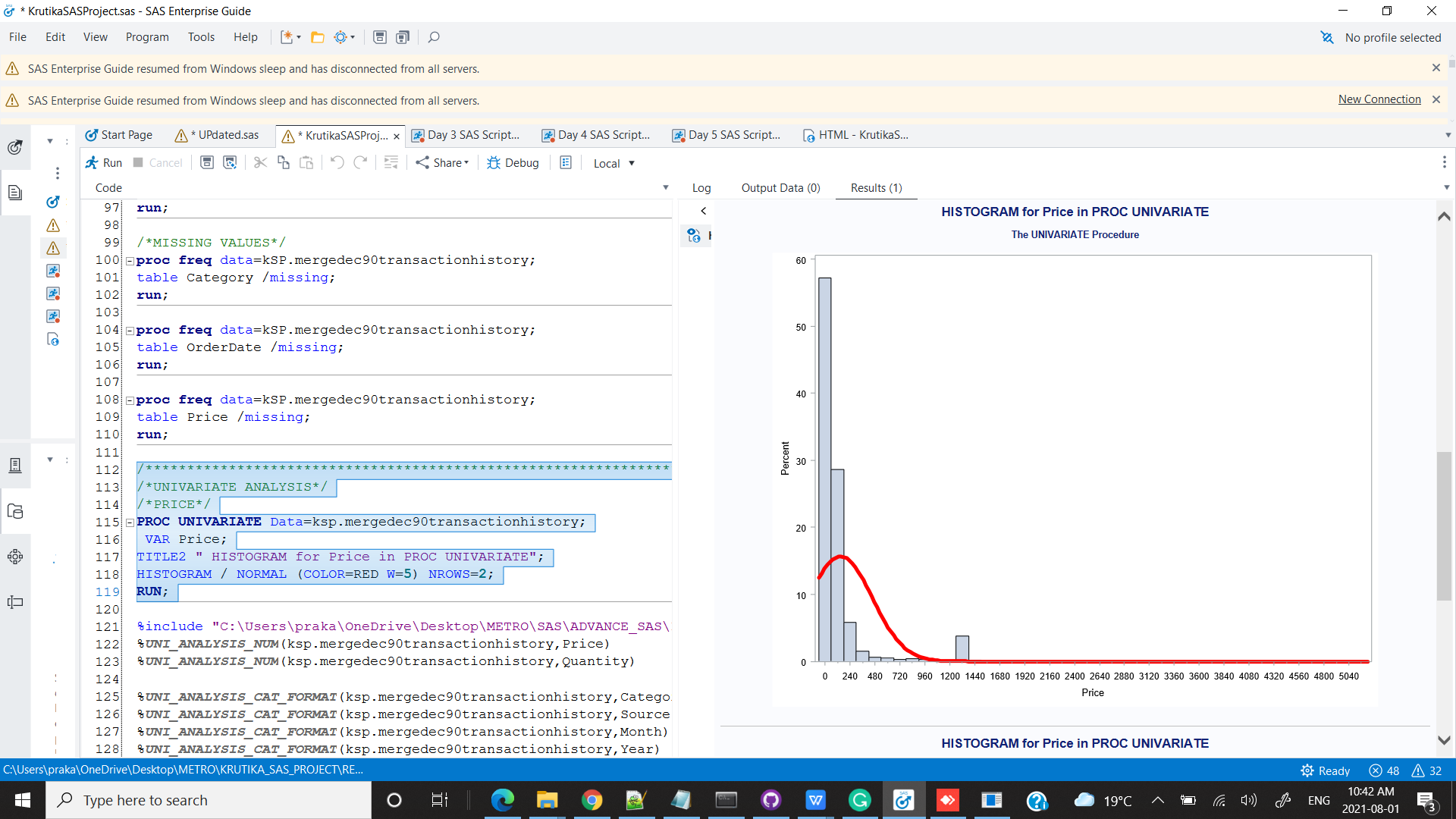
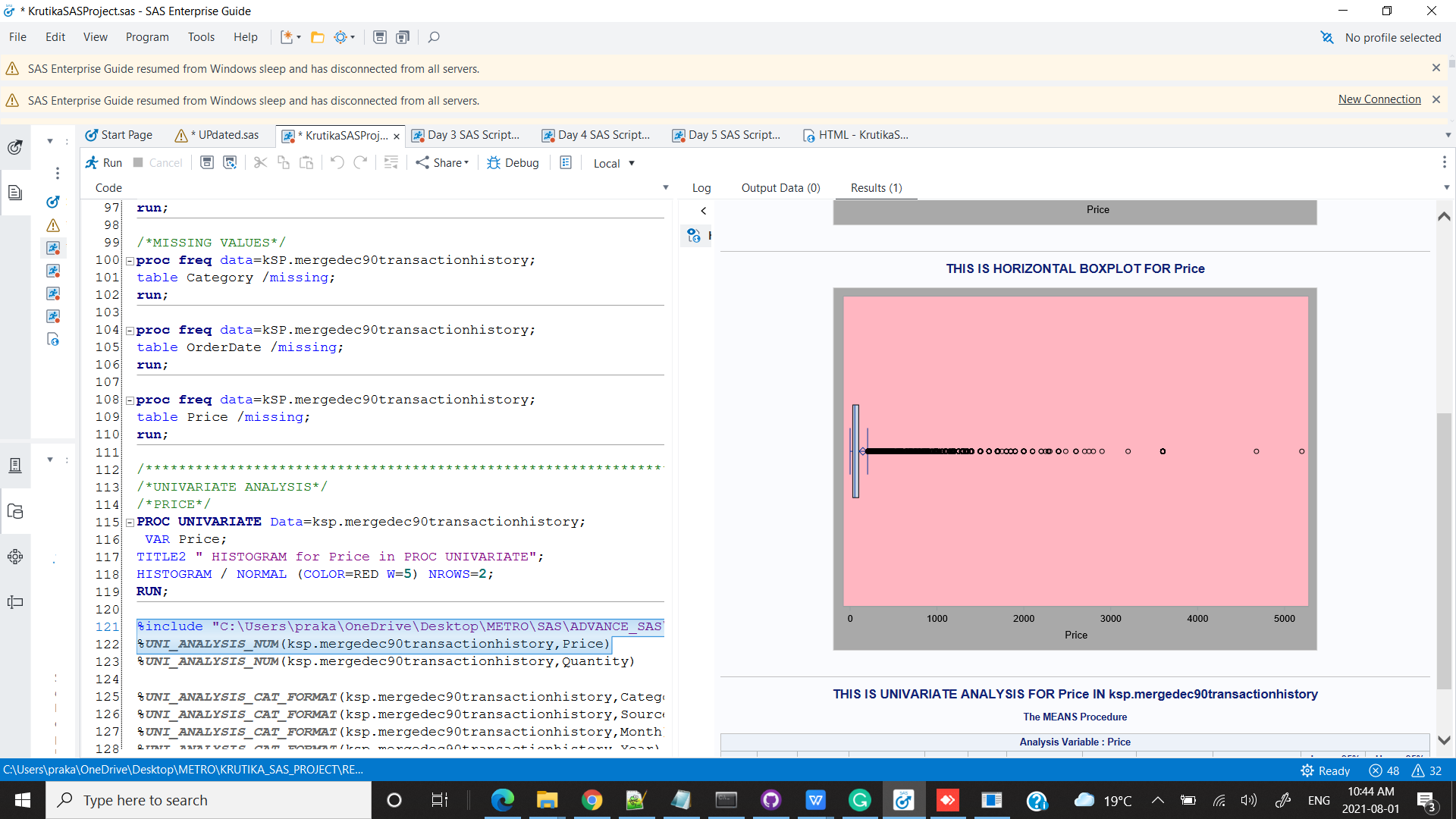
VAR Price;

TITLE2 " HISTOGRAM for Price in PROC UNIVARIATE";

HISTOGRAM / NORMAL (COLOR=RED W=**5**) NROWS=**2**;

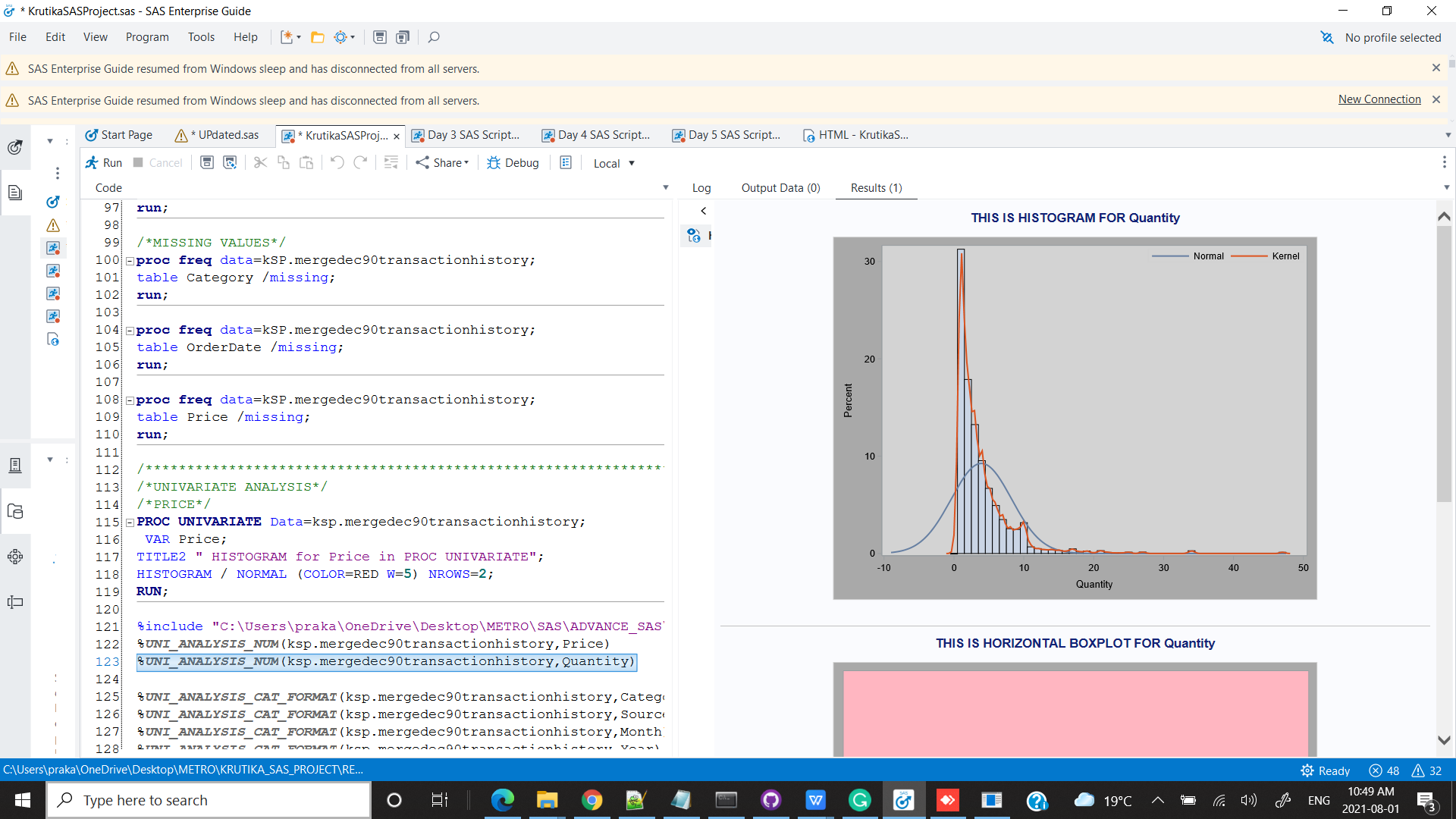
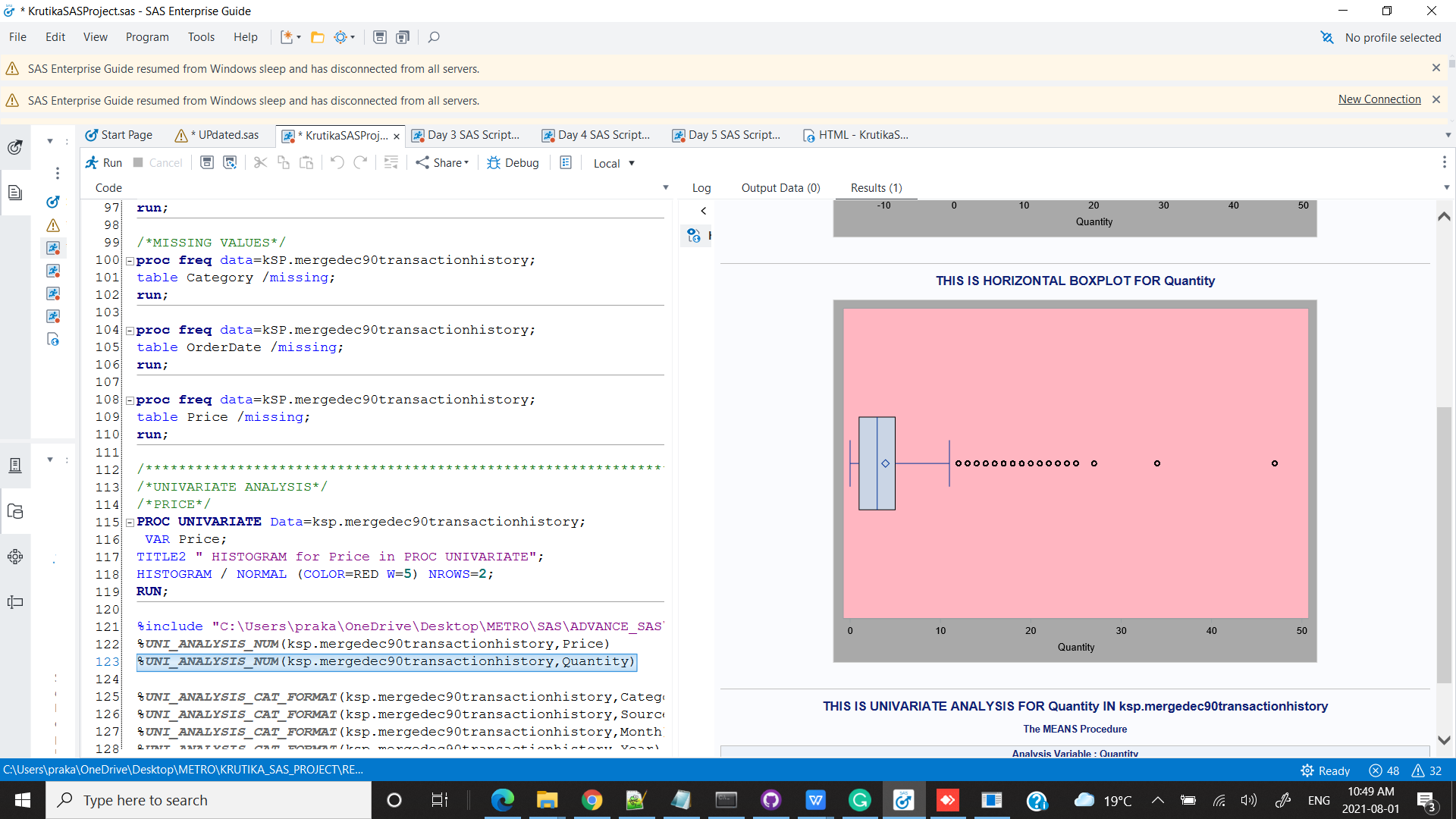
**RUN**;



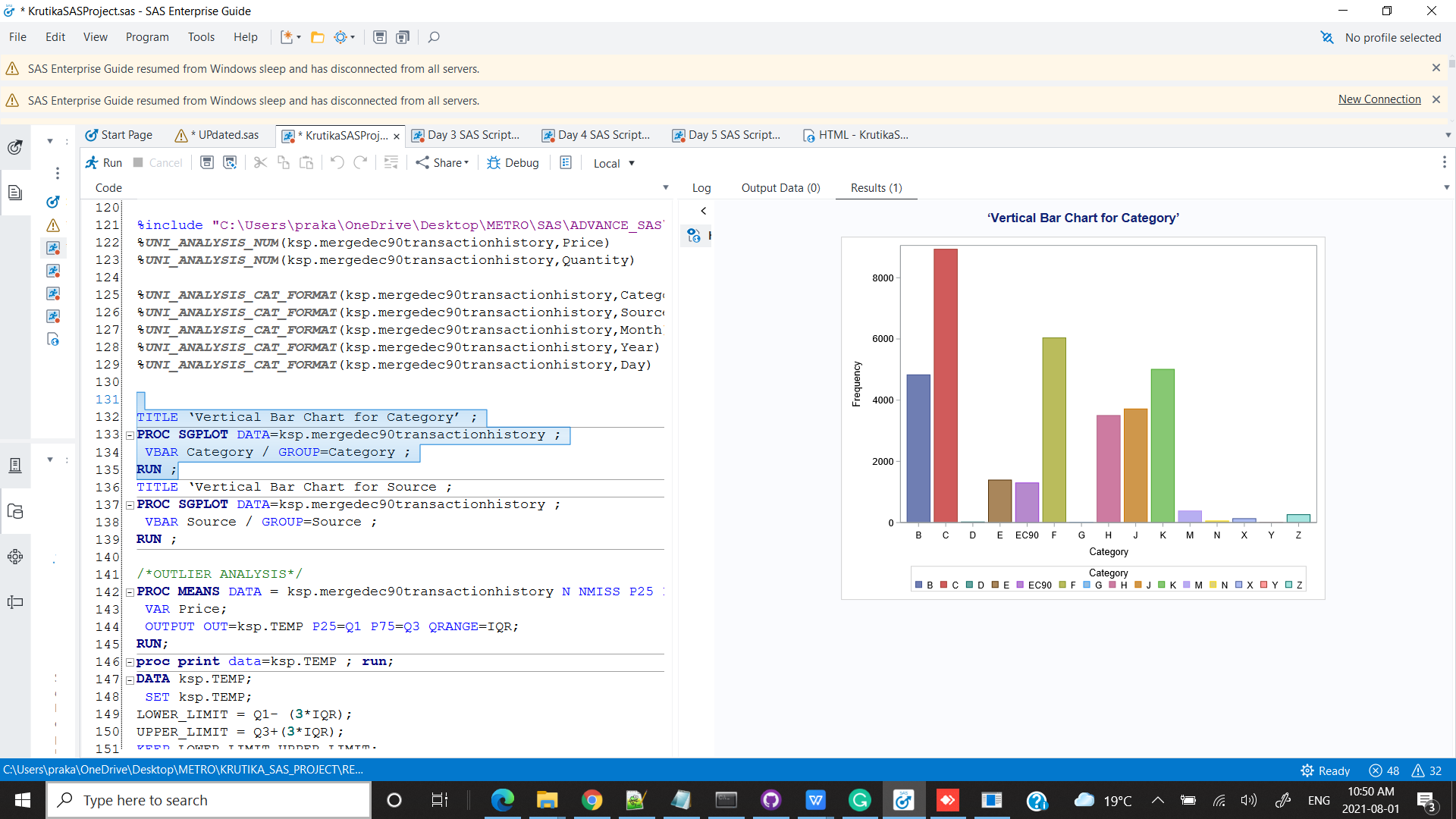
CONCLUSION : There are outliers and the data is positively skewed.

QUANTITY

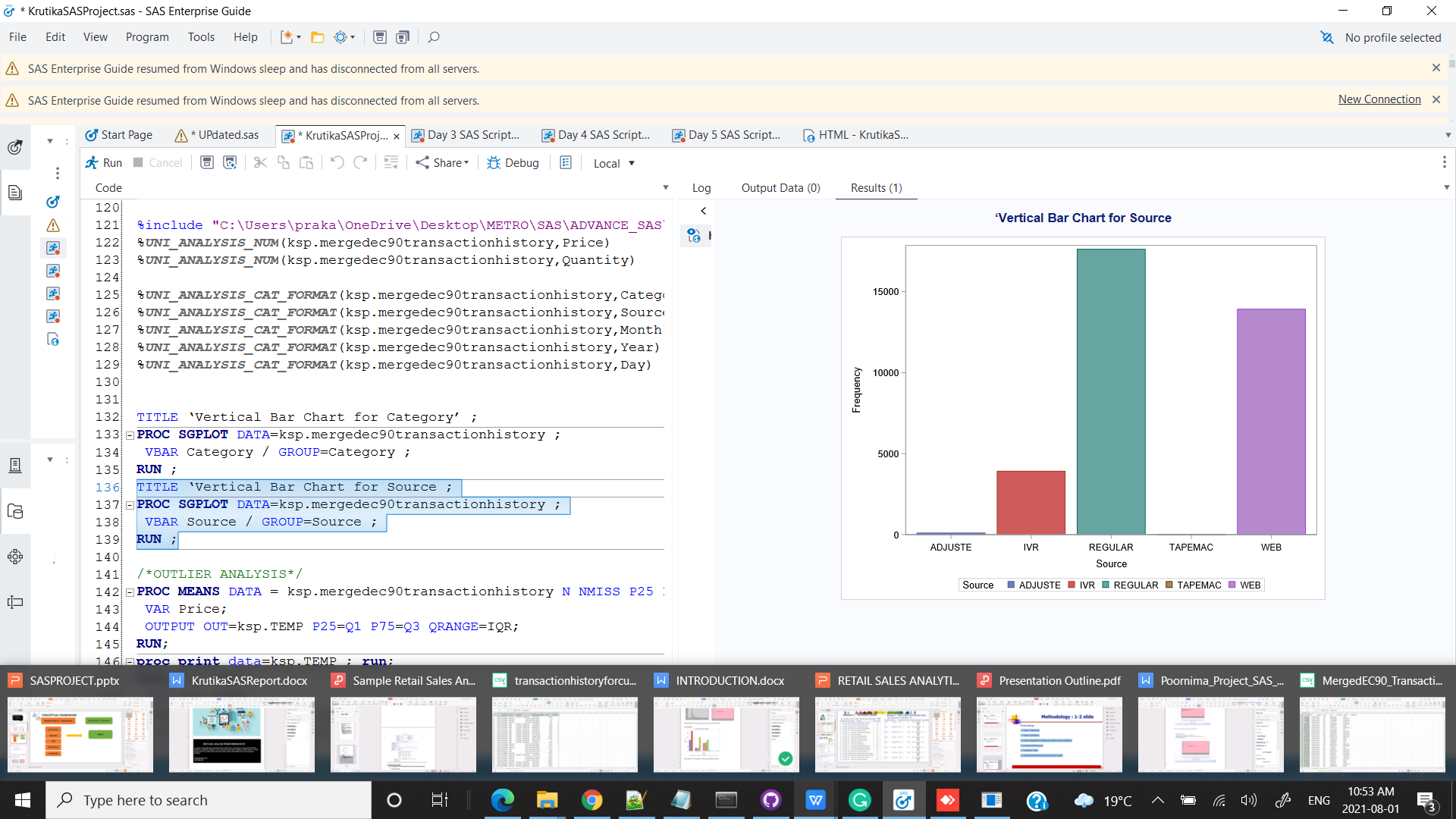
CONCLUSION : There are outliers and the data is positively skewed.

CATEGORYY



CONCLUSION : The categories C, F,K B has the highest frequency.

SOURCE



CONCLUSION : The categories Regular & Web has the highest frequency.

BIVARIATE ANALYSIS

/\*TOP 10 CUSTOMERID WITH HIGHEST SALES \*/

TITLE 'TOP 10 CUSTOMERS WITH HIGHEST RECORDS';

**PROC** **SQL** outobs=**10**;

create table ksp.Sales\_CustomerID as

SELECT CustomerID, SUM(Price) AS SALES

FROM ksp.mergedec90transactionhistory

group by CustomerID

ORDER BY SALES DESC;;

**QUIT**;

TITLE 'TOP 10 CUSTOMERS WITH HIGHEST RECORDS';

**proc** **sgplot** data=ksp.Sales\_CustomerID ;

hbar CustomerID / response= Sales

GROUP=CustomerID

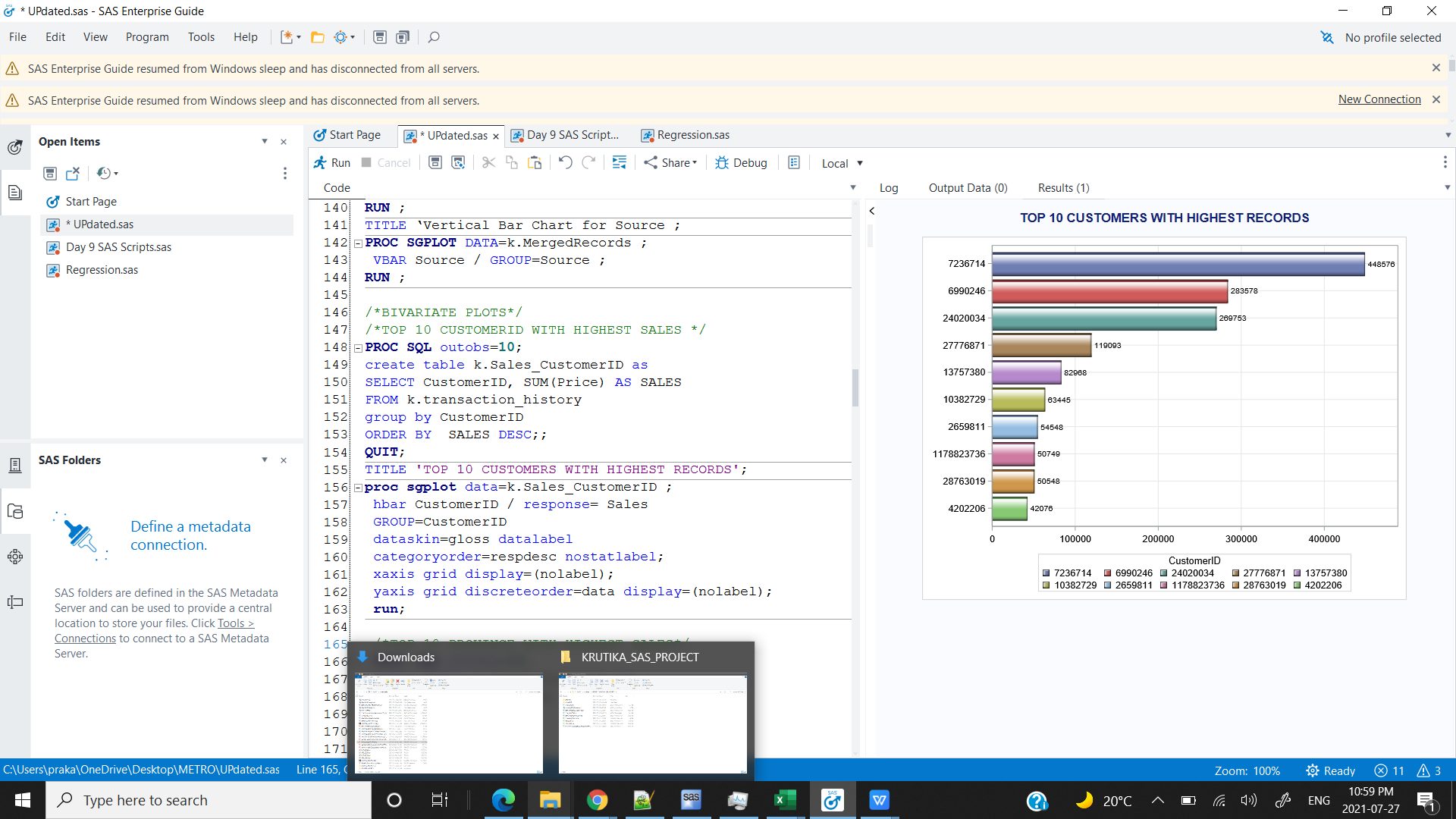
dataskin=gloss datalabel

categoryorder=respdesc nostatlabel;

xaxis grid display=(nolabel);

yaxis grid discreteorder=data display=(nolabel);

**run**;



/\*TOP 10 PROVINCE WITH HIGHEST SALES\*/

**PROC** **SQL** outobs=**10**;

create table ksp.Sales\_Province as

SELECT Prov, SUM(Price) AS SALES

FROM ksp.mergedec90transactionhistory

group by Prov

ORDER BY SALES DESC;;

**QUIT**;

title'TOP 10 PROVINCE WITH HIGHEST SALES';

**proc** **sgplot** data=ksp.Sales\_Province ;

hbar Prov / response= Sales

GROUP=Prov

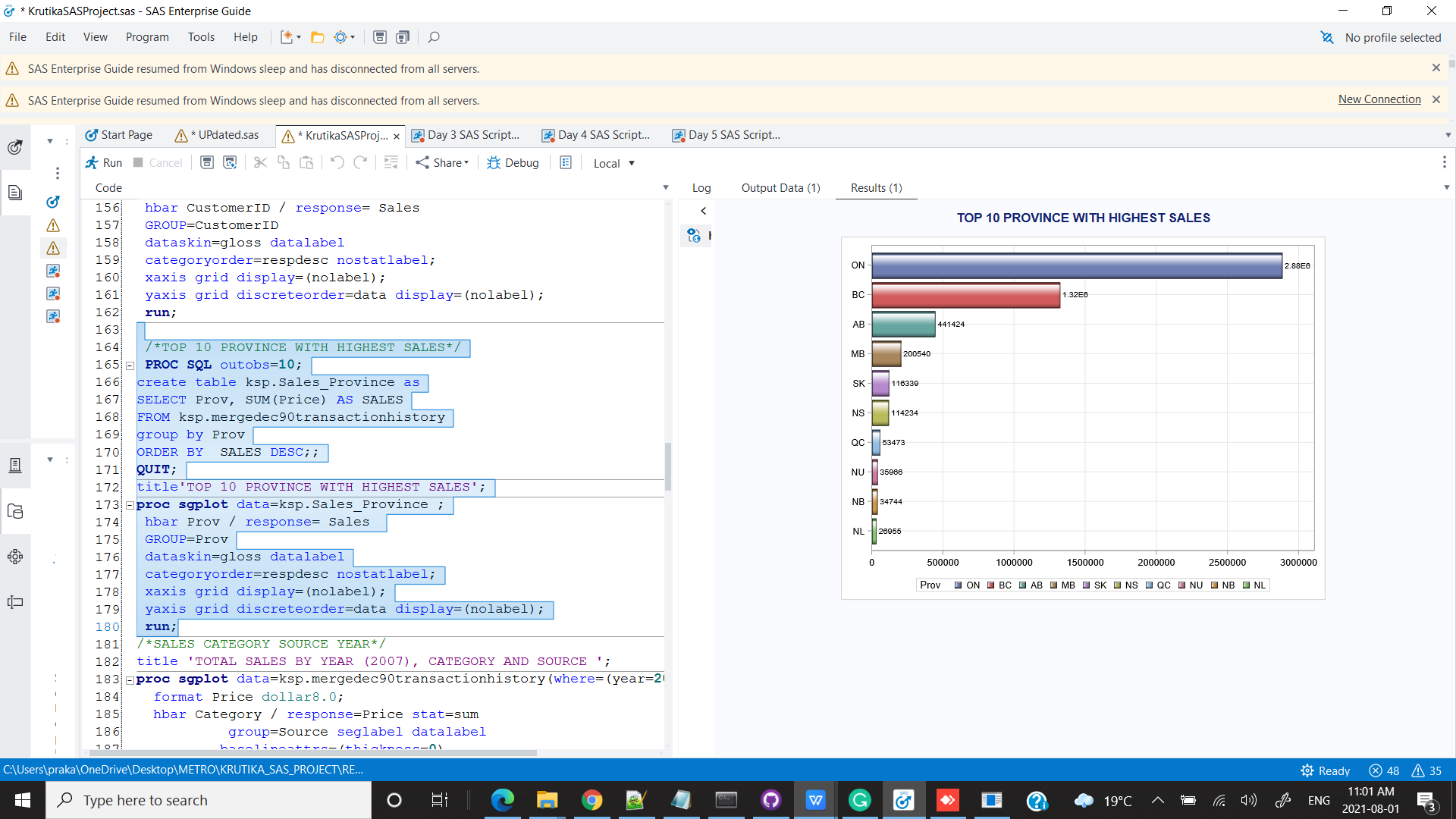
dataskin=gloss datalabel

categoryorder=respdesc nostatlabel;

xaxis grid display=(nolabel);

yaxis grid discreteorder=data display=(nolabel);

**run**;



CONCLUSION : Ontario has the highest frequency and Nunawat has the lowest Frequency

/\*TOP 10 PROVINCE WITH HIGHEST SALES\*/

**PROC** **SQL** outobs=**10**;

create table ksp.Sales\_Province as

SELECT Prov, SUM(Price) AS SALES

FROM ksp.mergedec90transactionhistory

group by Prov

ORDER BY SALES DESC;;

**QUIT**;

title'TOP 10 PROVINCE WITH HIGHEST SALES';

**proc** **sgplot** data=ksp.Sales\_Province ;

hbar Prov / response= Sales

GROUP=Prov

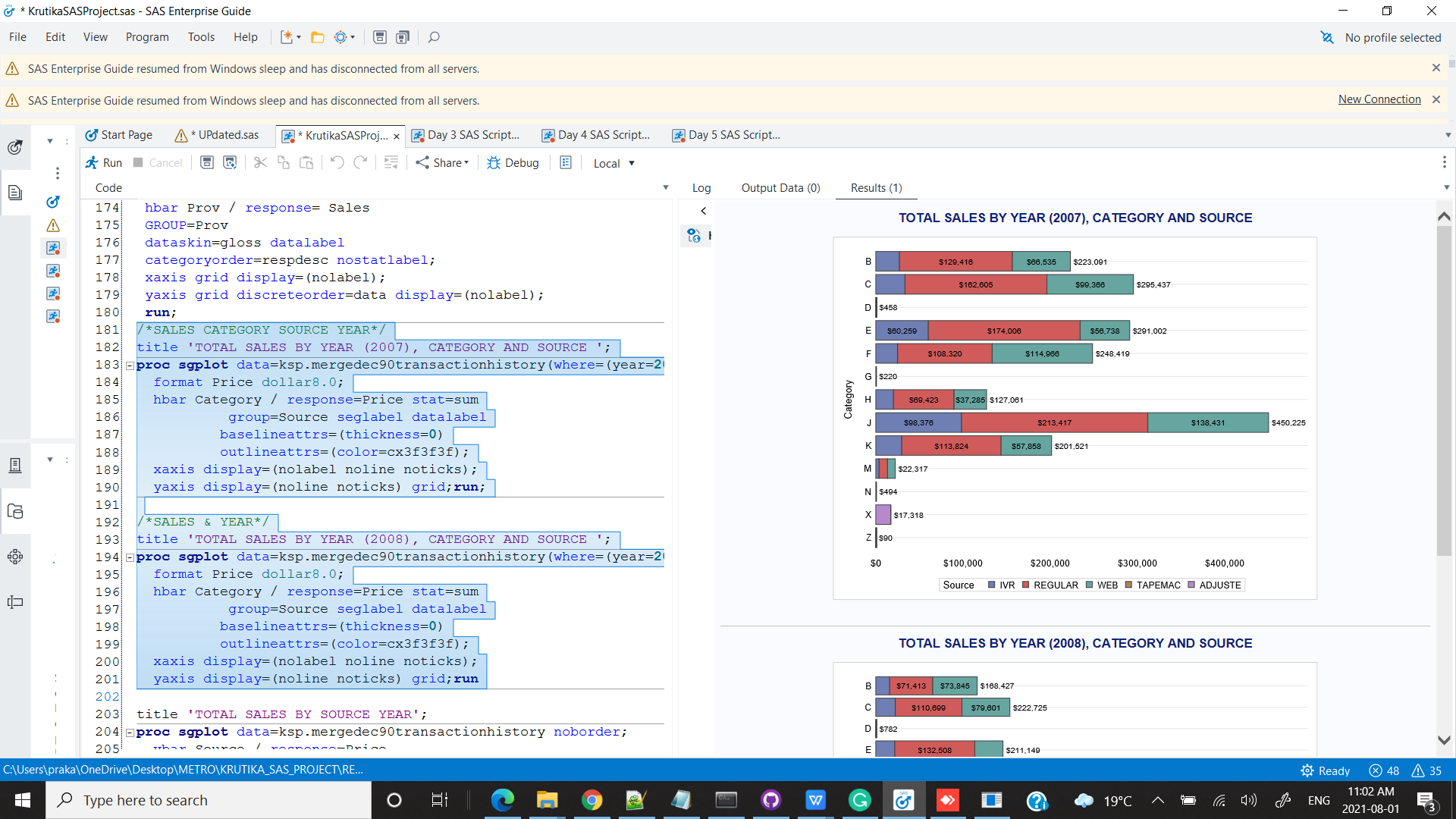
dataskin=gloss datalabel

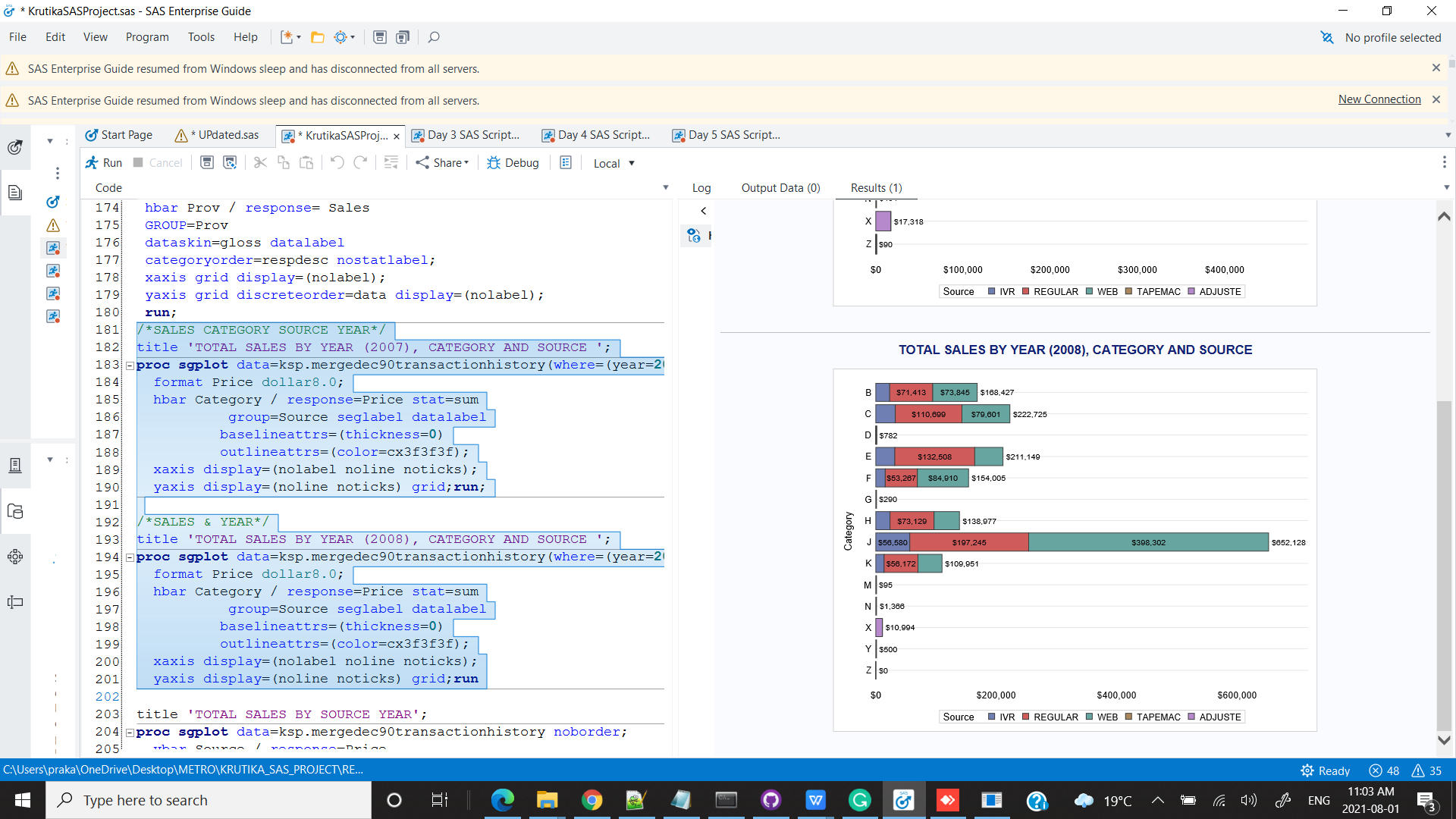
categoryorder=respdesc nostatlabel;

xaxis grid display=(nolabel);

yaxis grid discreteorder=data display=(nolabel);

**run**;





CONCLUSION : In 2007 Source REGULAR with Categories J,E& C had the highest sales & in 2008 WEB with categories J,E& C has the highest sales.

title 'TOTAL SALES BY SOURCE YEAR';

**proc** **sgplot** data=ksp.mergedec90transactionhistory noborder;

vbar Source / response=Price

group=Year groupdisplay=cluster

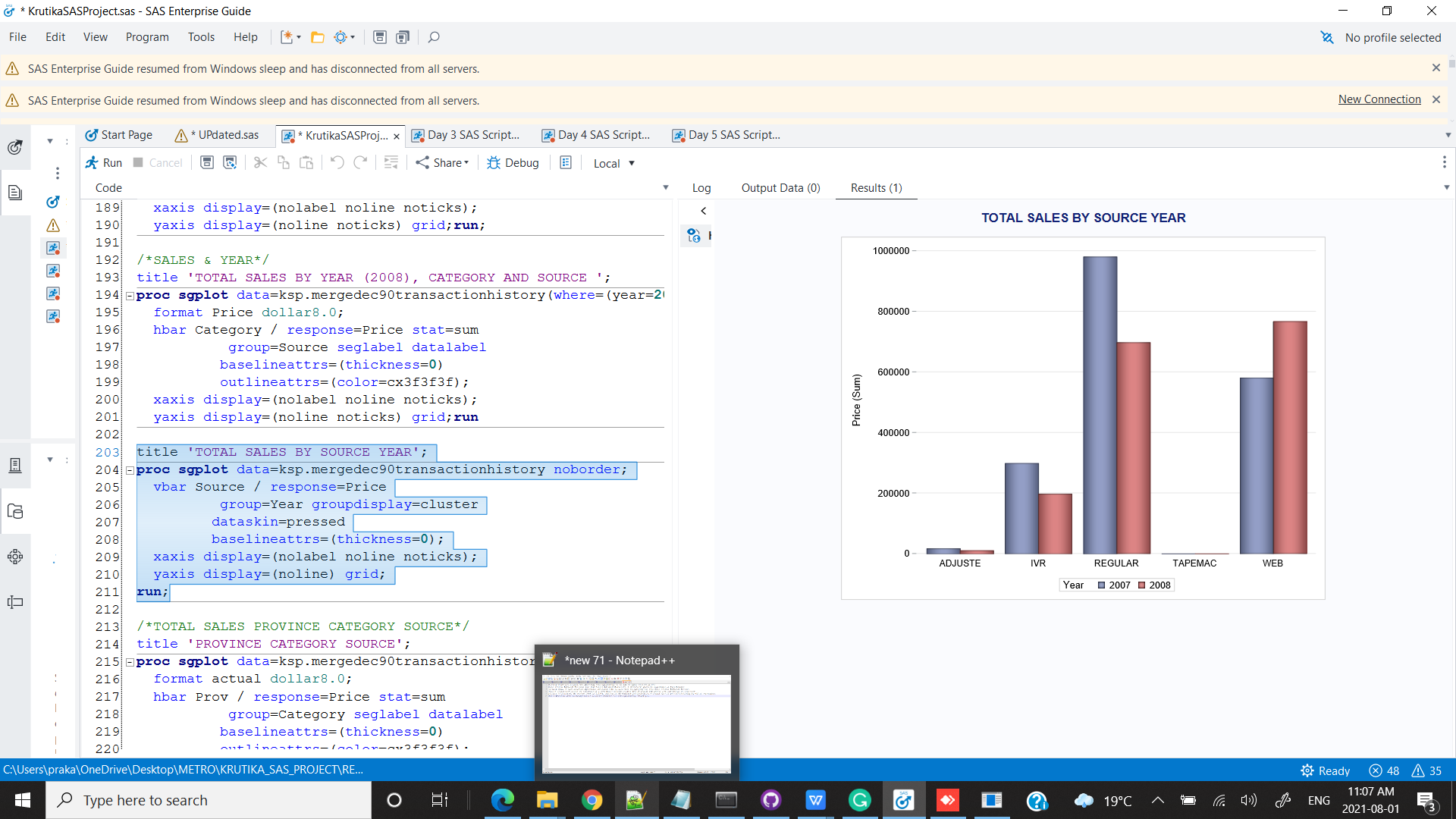
dataskin=pressed

baselineattrs=(thickness=**0**);

xaxis display=(nolabel noline noticks);

yaxis display=(noline) grid;

**run**;



CONCLUSION : In 2007 Source REGULAR had the highest sales & in 2008 WEB had the highest sales.

/\*TOTAL SALES PROVINCE CATEGORY SOURCE\*/

title 'PROVINCE CATEGORY SOURCE';

**proc** **sgplot** data=ksp.mergedec90transactionhistory noborder;

format actual dollar8.0;

hbar Prov / response=Price stat=sum

group=Category seglabel datalabel

baselineattrs=(thickness=**0**)

outlineattrs=(color=cx3f3f3f);

xaxis display=(nolabel noline noticks);

yaxis display=(noline noticks) grid;

**run**;



CONCLUSION : Ontario has the highest sales with category J , EC90 & F

MODELLING :

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*MODELLING\*/

Label Encoding the data

**data** ksp.var\_mlr;

set ksp.mergedec90transactionhistory;

drop OrderDate;

**run**;

**%macro** label\_encode(dataset,var);

proc sql noprint;

select distinct(&var)

into:val1-

from &dataset;

select count(distinct(&var)) into:mx from &dataset;

quit;

data ksp.encode\_city;

set &dataset;

%do i=**1** %to &mx;

if &var="&&&val&i" then encode\_city=&i;

%end;

run;

**%mend**;

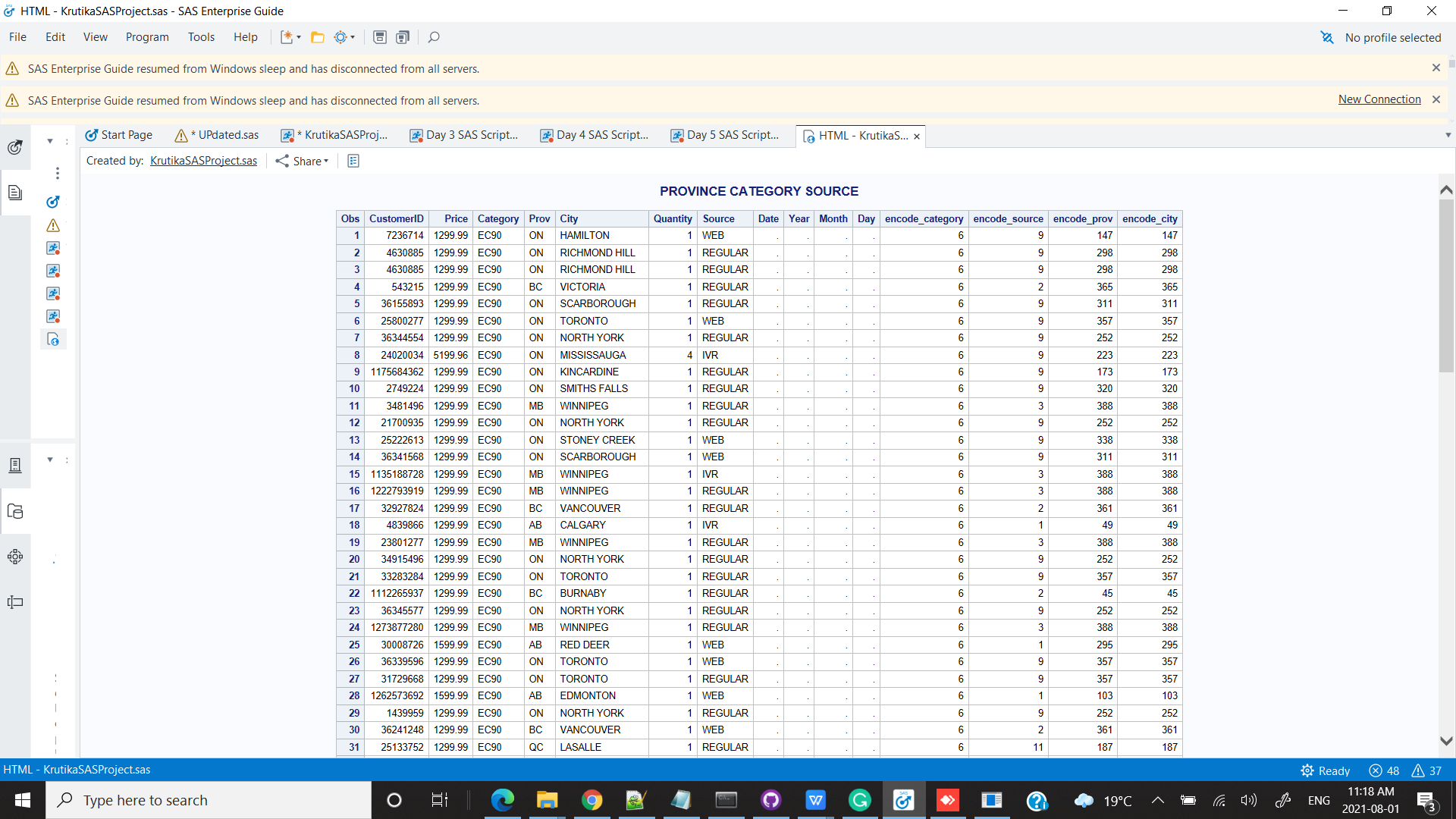
/\* define a macro to create dummy variables \*/

%***label\_encode***(ksp.var\_mlr,Category)

%***label\_encode***(ksp.encode\_category,Source)

%***label\_encode***(ksp.encode\_source,Prov)

%***label\_encode***(ksp.encode\_prov,City)



**proc** **print** data= ksp.encode\_city (obs=**100**);

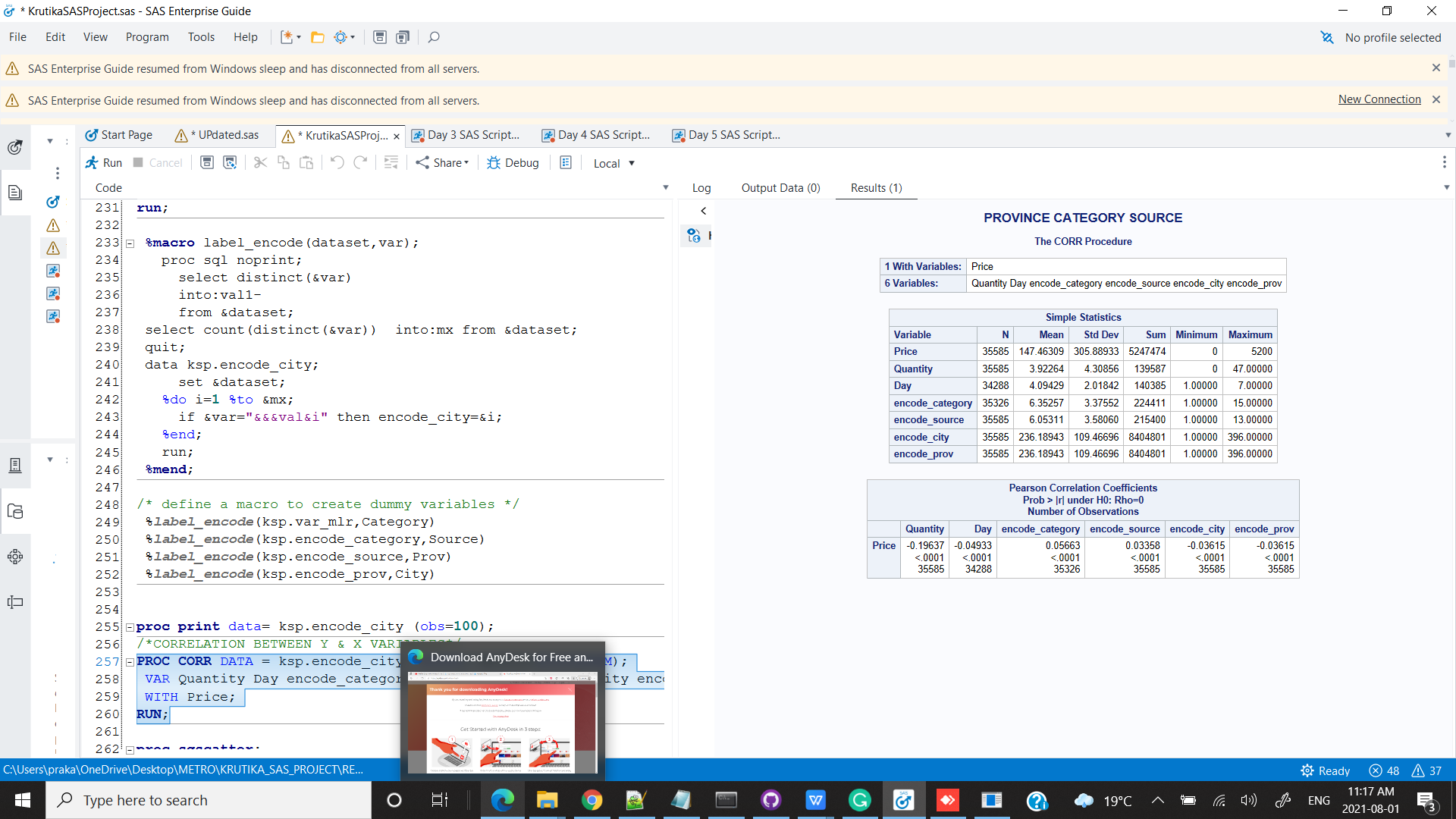
/\*CORRELATION BETWEEN Y & X VARIABLES\*/

**PROC** **CORR** DATA = ksp.encode\_city PLOTS= MATRIX(HISTOGRAM);

VAR Quantity Day encode\_category encode\_source encode\_city encode\_prov;

WITH Price;

**RUN**;

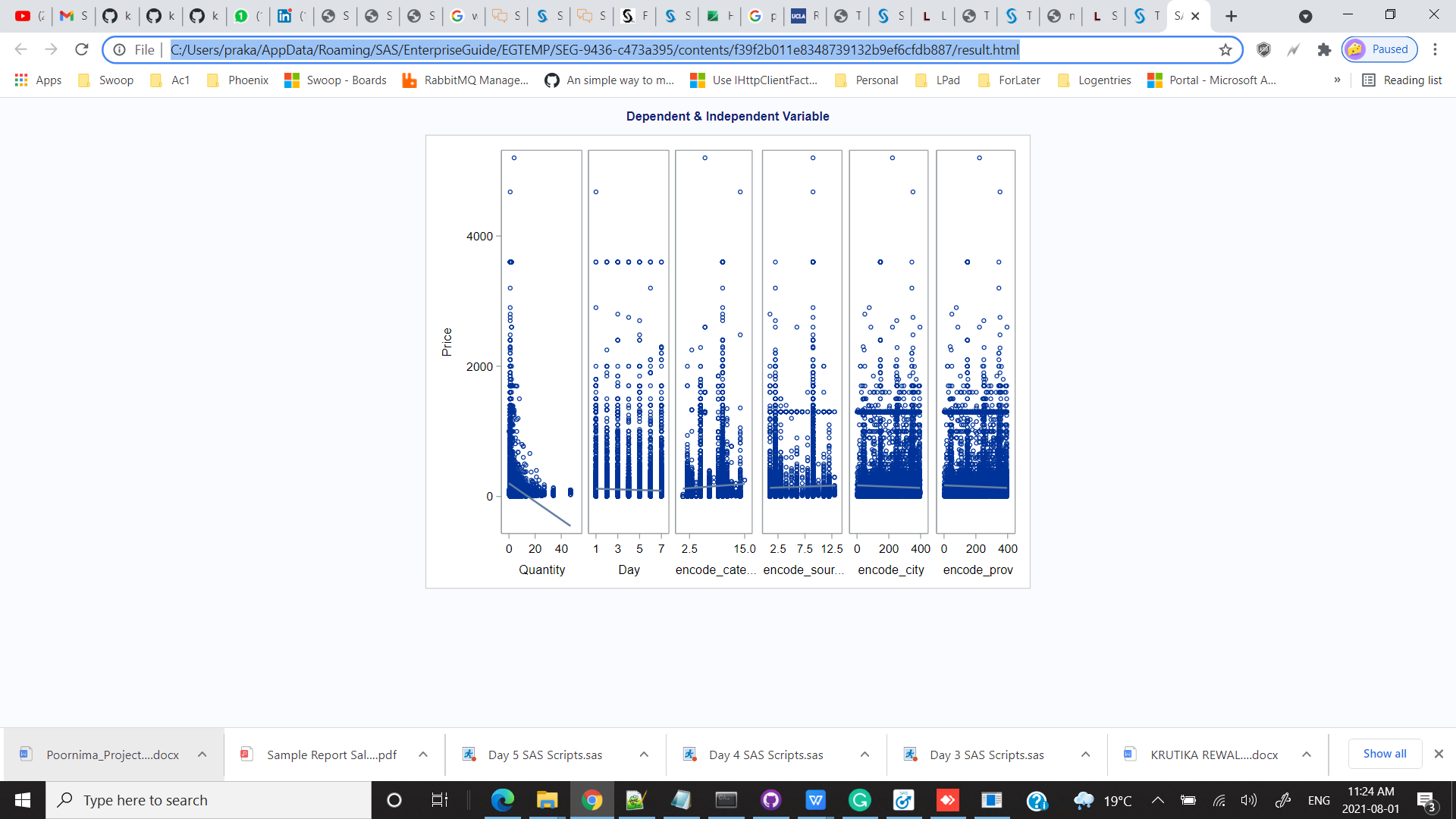


CONCLUSION : Category has the highest correlation with the PRice

**proc** **sgscatter**;

matrix Price Quantity Day encode\_category encode\_source encode\_city encode\_prov ;

**run**;

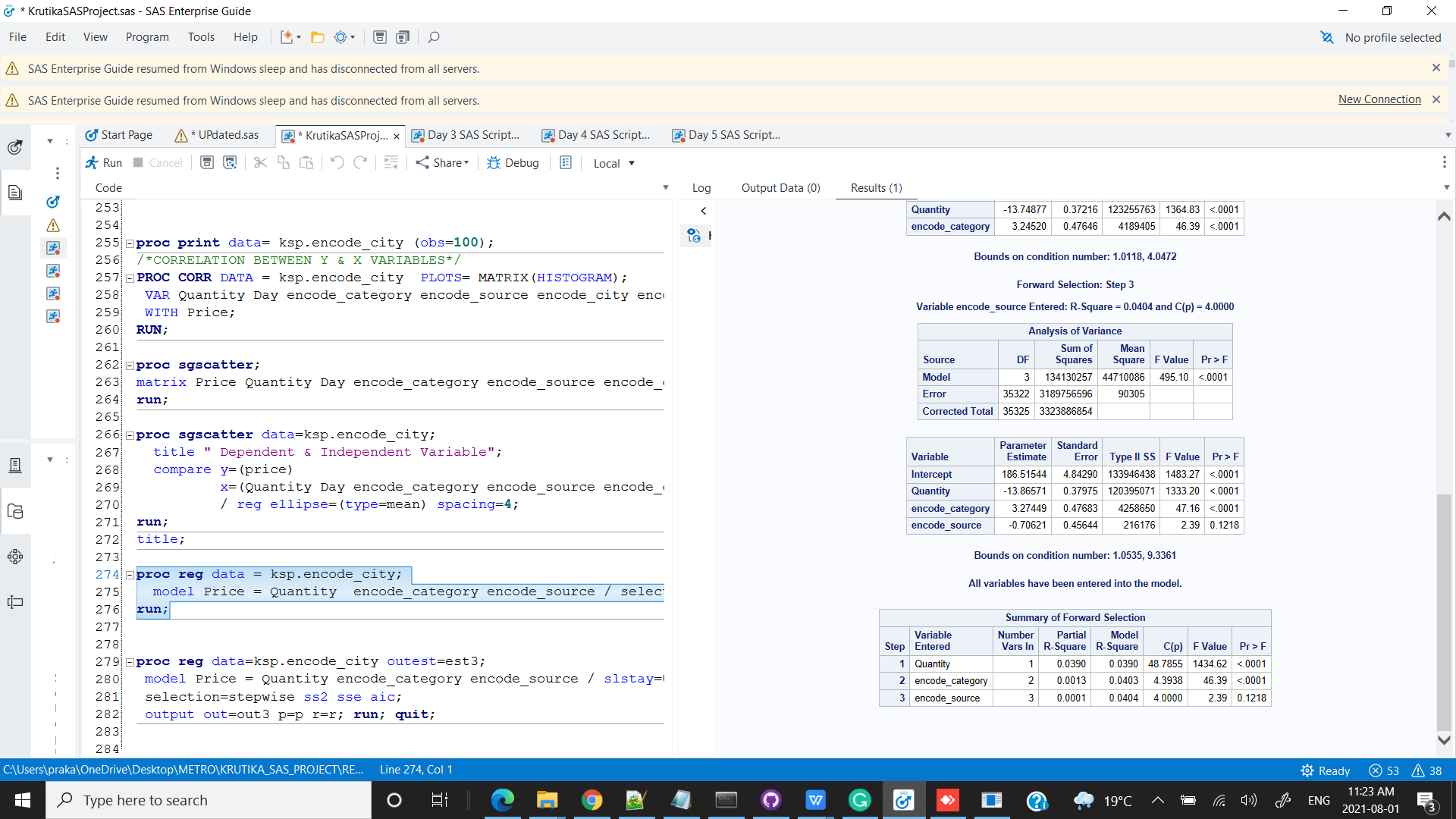


CONCLUSION : The scatter plot does not gives much insights on the correlation as all the independent variables are categorical and label encoded so there may be multi collinearity issues because of which we can’t see the clear rlationship between X & y variables.

**proc** **reg** data = ksp.encode\_city;

model Price = Quantity encode\_category encode\_source / selection = forward slentry = **0.99**;

**run**;



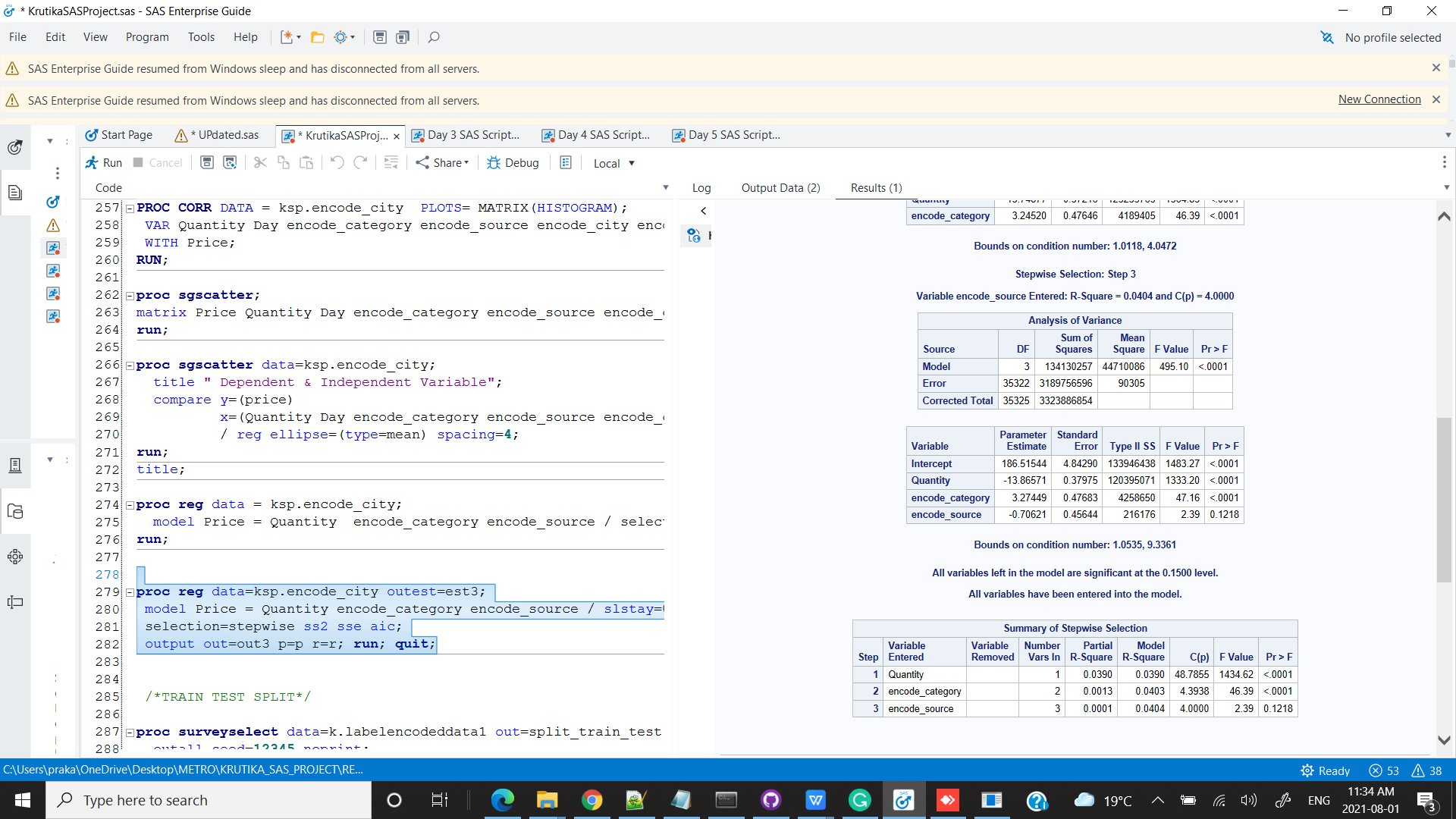
CONCLUSION : Adjusted R square is 0.0404 with Quantity , Category and Source as the variables rest other variables are excluded which are insignificant by forward selection

**proc** **reg** data=ksp.encode\_city outest=est3;

model Price = Quantity encode\_category encode\_source / slstay=**0.15** slentry=**0.15**

selection=stepwise ss2 sse aic;

output out=out3 p=p r=r; **run**; **quit**;



CONCLUSION : By using stepwise selection we get the same results for variables and R square.

/\*TRAIN TEST SPLIT\*/

**proc** **surveyselect** data=k.labelencodeddata1 out=split\_train\_test method=srs samprate=**0.70**

outall seed=**12345** noprint;

samplingunit encode\_customerid;

**run**;

/\*MODEL ON TRAIN DATA\*/

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\train\_data\_.csv'

out = ksp.train\_

dbms = csv

replace;

**run**;

**proc** **print** data= ksp.train\_(obs=**10**);

**run**;

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\test\_data\_.csv'

out = ksp.test\_

dbms = csv

replace;

**run**;

**proc** **print** data= ksp.test\_(obs=**10**);

**run**;

/\*Training the data\*/

**proc** **reg** data = ksp.train\_ noprint outest=estimates;

model Price = Quantity encode\_category encode\_source;

**run**;

/\*Predicted score values\*/

**proc** **score** data = ksp.test\_ score=estimates

out=scored type=parms;

var Quantity encode\_category encode\_source;

**run**;

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\UPDATED\test\_data\_.csv'

out = ksp.testtarget

dbms = csv

replace;

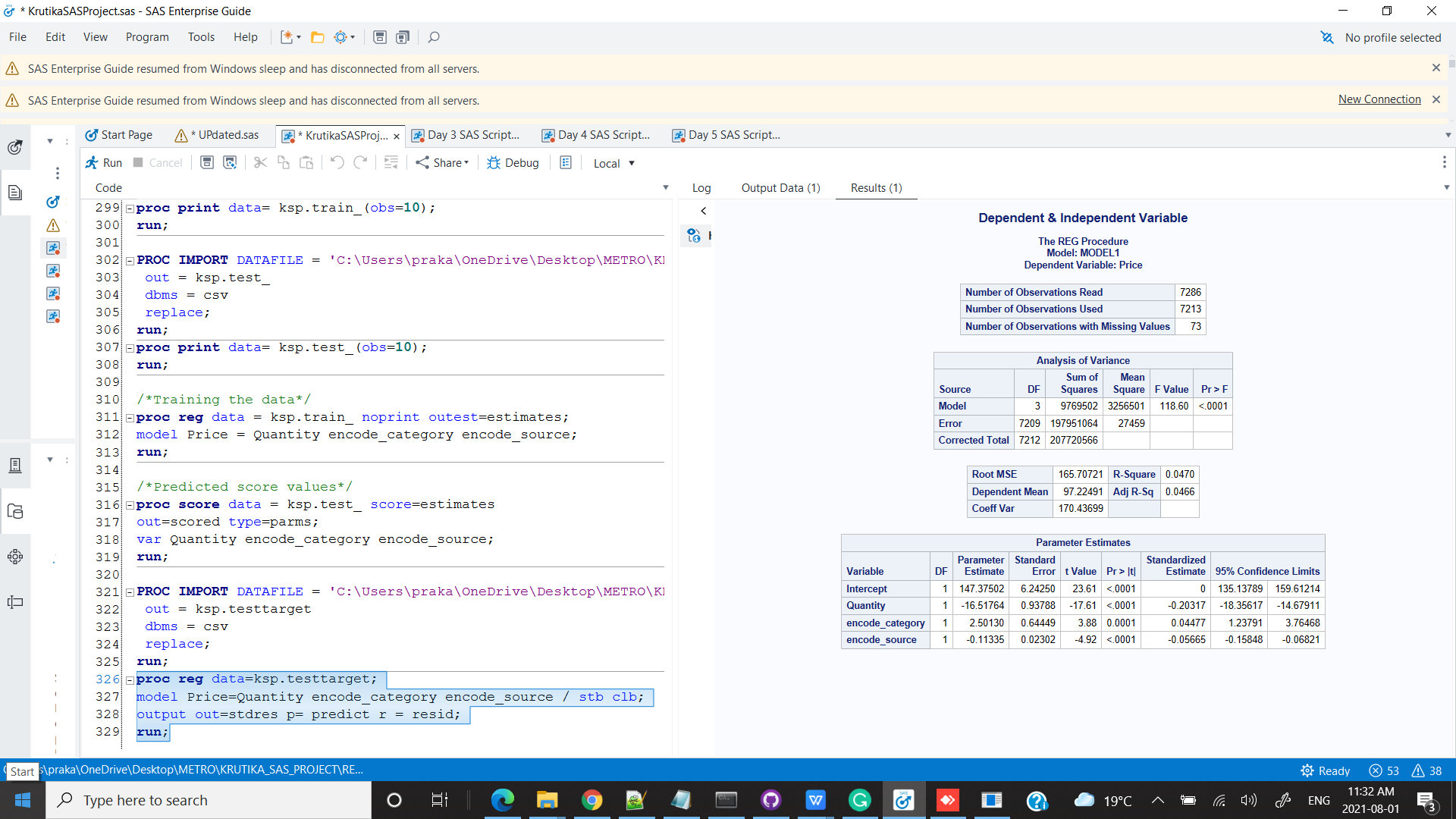
**run**;

**proc** **reg** data=ksp.testtarget;

model Price=Quantity encode\_category encode\_source / stb clb;

output out=stdres p= predict r = resid;

**run**;



CONCLUSION : For Predticting price the R square is bit high than train data with RMSE 165.7

/\*GLM SELECT\*/

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\traindata\_glmselect.csv'

out = ksp.data

dbms = csv

replace;

**run**;

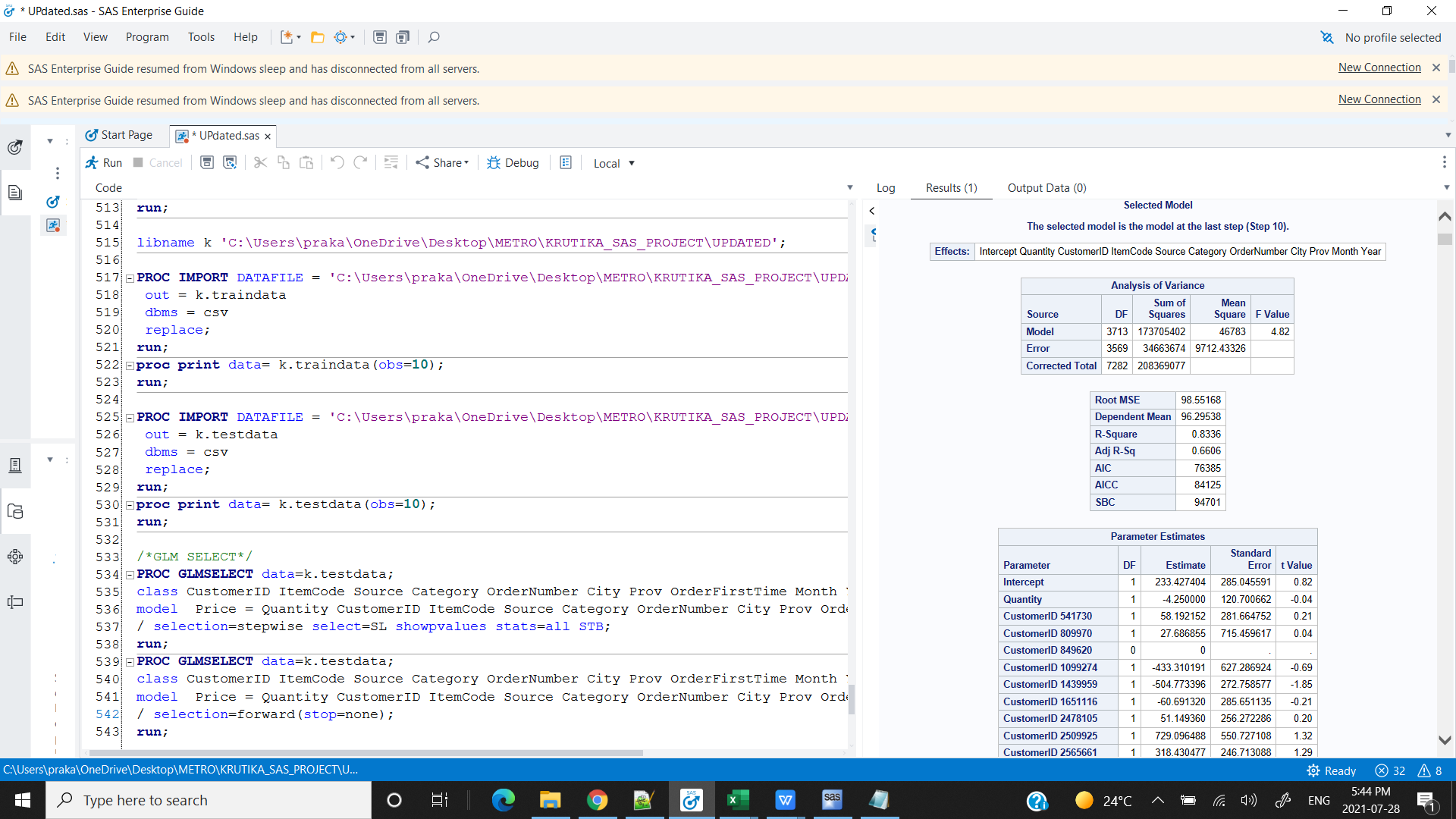
**PROC** **GLMSELECT** data=ksp.train\_;

class Quantity Source Category City Prov;

model Price = Quantity Source Category City Prov

/ selection=forward(stop=none);

**run**;



CONCLUSION : USing GLM SELECT Adjusted R square 0.66 .So GLM SELECT is the better approach for Price Prediction.

CONCLUSION :

1. Ontario has the highest sales and Nunavat has the lowest sales.
2. Ontario is buying more of Cosmetics and Jewellry by Web and Regular both the source.
3. Till 2007 Source with highest sales was Regular but in 2008 Web had the highest sales.
4. In categories Cosmetics had the highest sales followed by electronics.

RECCOMENDATIONS

1. Promote offers on highest buying product categories and also give a combo offers with the products with less sales.
2. Promote more advertising in other provinces where sales are low and try to pitch the products with good customer services.
3. Ontario is the leading province in all the product categories so there can be many opportunities for new business having some different or new products. We can expand more domestic stores in Ontario specially the city of Bellevile.

APPENDIX

/\*SETTING LIBRARY\*/

libname ksp 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT';

/\*IMPORTING ALL THE FILES\*/

/\*IMPORT FILES EC90\*/

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\ec90 data.csv'

out = ksp.ec90

dbms = csv

replace;

**run**;

/\*CHECK DATA WITH 100 OBSERVATIONS\*/

**proc** **print** data= ksp.ec90(obs=**100**);

**run**;

/\*IMPORT FILES TRANSACTION HISTORY\*/

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\transactionhistoryforcurrentcustomers.csv'

out = ksp.transaction\_history

dbms = csv

replace;

**run**;

/\*CHECK DATA WITH 100 OBSERVATIONS\*/

**proc** **print** data= ksp.transaction\_history(obs=**100**);

**run**;

/\*IMPORT FILES MASTER TABLE CUSTOMERID & LOCATION\*/

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\MASTERTABLE\_CUSTOMERID\_LOCATION.csv'

out = ksp.mastertable

dbms = csv

replace;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*JOIN EC90 & TRANSACTION HISTORY WITH MASTER TABLE\*/

**proc** **sql**;

create table ksp.mergedec90transactionhistory as

select CustomerID,Price,Category,Prov,City,Quantity,Source

from ksp.ec90 as a

union all

select a.CustomerID,a.Price,Category,b.Prov,b.City,a.Quantity,Source,a.OrderDate

from ksp.transaction\_history as a join ksp.mastertable as b on

a.CustomerID = b.CustomerID;

**quit**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*ADDING COLUMNS BASED ON DATE\*/

/\*ORDERDATE\*/

**data** ksp.mergedec90transactionhistory;

set ksp.mergedec90transactionhistory;

Date = datepart(OrderDate);

format Date date9.;

**run**;

/\*CHECK DATA WITH 100 OBSERVATIONS\*/

**proc** **print** data = ksp.mergedec90transactionhistory(obs=**100**);**run**;

**quit**;

/\*ORDERDATE YEAR\*/

**data** ksp.mergedec90transactionhistory;

set ksp.mergedec90transactionhistory;

Year = year(Date);

**run**;

**proc** **print** data = ksp.mergedec90transactionhistory;**run**;

/\*ORDERDATE MONTH\*/

**data** ksp.mergedec90transactionhistory;

set ksp.mergedec90transactionhistory;

Month = intnx('month', Date, **0**);

format Month monname3.;

**run**;

**proc** **print** data = ksp.mergedec90transactionhistory(obs=**100**);**run**;

/\*ORDERDATE DAY\*/

**data** ksp.mergedec90transactionhistory;

set ksp.mergedec90transactionhistory;

Day = weekday(Date);

**run**;

/\*CHECK DATA WITH 100 OBSERVATIONS\*/

**proc** **print** data = ksp.mergedec90transactionhistory(obs=**2000**);**run**;

/\*EXPORT TO EXCEL\*/

**proc** **export**

data=ksp.mergedec90transactionhistory

dbms=xlsx

outfile="C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\MERGERECORDSEC90TRANSACTIONHISTORY.xlsx"

replace;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*TOTAL NO OF RECORDS\*/

**proc** **contents** data=ksp.mergedec90transactionhistory varnum;

**run**;

/\*MISSING VALUES\*/

**proc** **freq** data=kSP.mergedec90transactionhistory;

table Category /missing;

**run**;

**proc** **freq** data=kSP.mergedec90transactionhistory;

table OrderDate /missing;

**run**;

**proc** **freq** data=kSP.mergedec90transactionhistory;

table Price /missing;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*UNIVARIATE ANALYSIS\*/

/\*PRICE\*/

**PROC** **UNIVARIATE** Data=ksp.mergedec90transactionhistory;

VAR Price;

TITLE2 " HISTOGRAM for Price in PROC UNIVARIATE";

HISTOGRAM / NORMAL (COLOR=RED W=**5**) NROWS=**2**;

**RUN**;

%include "C:\Users\praka\OneDrive\Desktop\METRO\SAS\ADVANCE\_SAS\PROJECT\KRUTIKA\_PROJECT\univariate\_analysis.sas";

%***UNI\_ANALYSIS\_NUM***(ksp.mergedec90transactionhistory,Price)

%***UNI\_ANALYSIS\_NUM***(ksp.mergedec90transactionhistory,Quantity)

%***UNI\_ANALYSIS\_CAT\_FORMAT***(ksp.mergedec90transactionhistory,Category)

%***UNI\_ANALYSIS\_CAT\_FORMAT***(ksp.mergedec90transactionhistory,Source)

%***UNI\_ANALYSIS\_CAT\_FORMAT***(ksp.mergedec90transactionhistory,Month)

%***UNI\_ANALYSIS\_CAT\_FORMAT***(ksp.mergedec90transactionhistory,Year)

%***UNI\_ANALYSIS\_CAT\_FORMAT***(ksp.mergedec90transactionhistory,Day)

TITLE ‘Vertical Bar Chart for Category’ ;

**PROC** **SGPLOT** DATA=ksp.mergedec90transactionhistory ;

VBAR Category / GROUP=Category ;

**RUN** ;

TITLE ‘Vertical Bar Chart for Source ;

**PROC** **SGPLOT** DATA=ksp.mergedec90transactionhistory ;

VBAR Source / GROUP=Source ;

**RUN** ;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*BIVARIATE ANALYSIS\*/

/\*TOP 10 CUSTOMERID WITH HIGHEST SALES \*/

TITLE 'TOP 10 CUSTOMERS WITH HIGHEST RECORDS';

**PROC** **SQL** outobs=**10**;

create table ksp.Sales\_CustomerID as

SELECT CustomerID, SUM(Price) AS SALES

FROM ksp.mergedec90transactionhistory

group by CustomerID

ORDER BY SALES DESC;;

**QUIT**;

TITLE 'TOP 10 CUSTOMERS WITH HIGHEST RECORDS';

**proc** **sgplot** data=ksp.Sales\_CustomerID ;

hbar CustomerID / response= Sales

GROUP=CustomerID

dataskin=gloss datalabel

categoryorder=respdesc nostatlabel;

xaxis grid display=(nolabel);

yaxis grid discreteorder=data display=(nolabel);

**run**;

/\*TOP 10 PROVINCE WITH HIGHEST SALES\*/

**PROC** **SQL** outobs=**10**;

create table ksp.Sales\_Province as

SELECT Prov, SUM(Price) AS SALES

FROM ksp.mergedec90transactionhistory

group by Prov

ORDER BY SALES DESC;;

**QUIT**;

title'TOP 10 PROVINCE WITH HIGHEST SALES';

**proc** **sgplot** data=ksp.Sales\_Province ;

hbar Prov / response= Sales

GROUP=Prov

dataskin=gloss datalabel

categoryorder=respdesc nostatlabel;

xaxis grid display=(nolabel);

yaxis grid discreteorder=data display=(nolabel);

**run**;

/\*SALES CATEGORY SOURCE YEAR\*/

title 'TOTAL SALES BY YEAR (2007), CATEGORY AND SOURCE ';

**proc** **sgplot** data=ksp.mergedec90transactionhistory(where=(year=**2007**))noborder;

format Price dollar8.0;

hbar Category / response=Price stat=sum

group=Source seglabel datalabel

baselineattrs=(thickness=**0**)

outlineattrs=(color=cx3f3f3f);

xaxis display=(nolabel noline noticks);

yaxis display=(noline noticks) grid;**run**;

/\*SALES & YEAR\*/

title 'TOTAL SALES BY YEAR (2008), CATEGORY AND SOURCE ';

**proc** **sgplot** data=ksp.mergedec90transactionhistory(where=(year=**2008**))noborder;

format Price dollar8.0;

hbar Category / response=Price stat=sum

group=Source seglabel datalabel

baselineattrs=(thickness=**0**)

outlineattrs=(color=cx3f3f3f);

xaxis display=(nolabel noline noticks);

yaxis display=(noline noticks) grid;**run**

title 'TOTAL SALES BY SOURCE YEAR';

**proc** **sgplot** data=ksp.mergedec90transactionhistory noborder;

vbar Source / response=Price

group=Year groupdisplay=cluster

dataskin=pressed

baselineattrs=(thickness=**0**);

xaxis display=(nolabel noline noticks);

yaxis display=(noline) grid;

**run**;

/\*TOTAL SALES PROVINCE CATEGORY SOURCE\*/

title 'PROVINCE CATEGORY SOURCE';

**proc** **sgplot** data=ksp.mergedec90transactionhistory noborder;

format actual dollar8.0;

hbar Prov / response=Price stat=sum

group=Category seglabel datalabel

baselineattrs=(thickness=**0**)

outlineattrs=(color=cx3f3f3f);

xaxis display=(nolabel noline noticks);

yaxis display=(noline noticks) grid;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*MODELLING\*/

**data** ksp.var\_mlr;

set ksp.mergedec90transactionhistory;

drop OrderDate;

**run**;

**%macro** label\_encode(dataset,var);

proc sql noprint;

select distinct(&var)

into:val1-

from &dataset;

select count(distinct(&var)) into:mx from &dataset;

quit;

data ksp.encode\_city;

set &dataset;

%do i=**1** %to &mx;

if &var="&&&val&i" then encode\_city=&i;

%end;

run;

**%mend**;

/\* define a macro to create dummy variables \*/

%***label\_encode***(ksp.var\_mlr,Category)

%***label\_encode***(ksp.encode\_category,Source)

%***label\_encode***(ksp.encode\_source,Prov)

%***label\_encode***(ksp.encode\_prov,City)

**proc** **print** data= ksp.encode\_city (obs=**100**);

/\*CORRELATION BETWEEN Y & X VARIABLES\*/

**PROC** **CORR** DATA = ksp.encode\_city PLOTS= MATRIX(HISTOGRAM);

VAR Quantity Day encode\_category encode\_source encode\_city encode\_prov;

WITH Price;

**RUN**;

**proc** **sgscatter**;

matrix Price Quantity Day encode\_category encode\_source encode\_city encode\_prov ;

**run**;

**proc** **sgscatter** data=ksp.encode\_city;

title " Dependent & Independent Variable";

compare y=(price)

x=(Quantity Day encode\_category encode\_source encode\_city encode\_prov )

/ reg ellipse=(type=mean) spacing=**4**;

**run**;

title;

**proc** **reg** data = ksp.encode\_city;

model Price = Quantity encode\_category encode\_source / selection = forward slentry = **0.99**;

**run**;

**proc** **reg** data=ksp.encode\_city outest=est3;

model Price = Quantity encode\_category encode\_source / slstay=**0.15** slentry=**0.15**

selection=stepwise ss2 sse aic;

output out=out3 p=p r=r; **run**; **quit**;

/\*TRAIN TEST SPLIT\*/

**proc** **surveyselect** data=k.labelencodeddata1 out=split\_train\_test method=srs samprate=**0.70**

outall seed=**12345** noprint;

samplingunit encode\_customerid;

**run**;

/\*MODEL ON TRAIN DATA\*/

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\train\_data\_.csv'

out = ksp.train\_

dbms = csv

replace;

**run**;

**proc** **print** data= ksp.train\_(obs=**10**);

**run**;

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\test\_data\_.csv'

out = ksp.test\_

dbms = csv

replace;

**run**;

**proc** **print** data= ksp.test\_(obs=**10**);

**run**;

/\*Training the data\*/

**proc** **reg** data = ksp.train\_ noprint outest=estimates;

model Price = Quantity encode\_category encode\_source;

**run**;

/\*Predicted score values\*/

**proc** **score** data = ksp.test\_ score=estimates

out=scored type=parms;

var Quantity encode\_category encode\_source;

**run**;

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\UPDATED\test\_data\_.csv'

out = ksp.testtarget

dbms = csv

replace;

**run**;

**proc** **reg** data=ksp.testtarget;

model Price=Quantity encode\_category encode\_source / stb clb;

output out=stdres p= predict r = resid;

**run**;

/\*GLM SELECT\*/

**PROC** **IMPORT** DATAFILE = 'C:\Users\praka\OneDrive\Desktop\METRO\KRUTIKA\_SAS\_PROJECT\REPORT\traindata\_glmselect.csv'

out = ksp.data

dbms = csv

replace;

**run**;

**PROC** **GLMSELECT** data=ksp.train\_;

class Quantity Source Category City Prov;

model Price = Quantity Source Category City Prov

/ selection=forward(stop=none);

**run**;