SAS Portfolio Project – Analysis of Orders

Anthony Vuolo

MIS500 – Foundations of Data Analytics

Colorado State University-Global Campus

Dr. Davis

6/07/2020

SAS Portfolio Project – Analysis of Orders

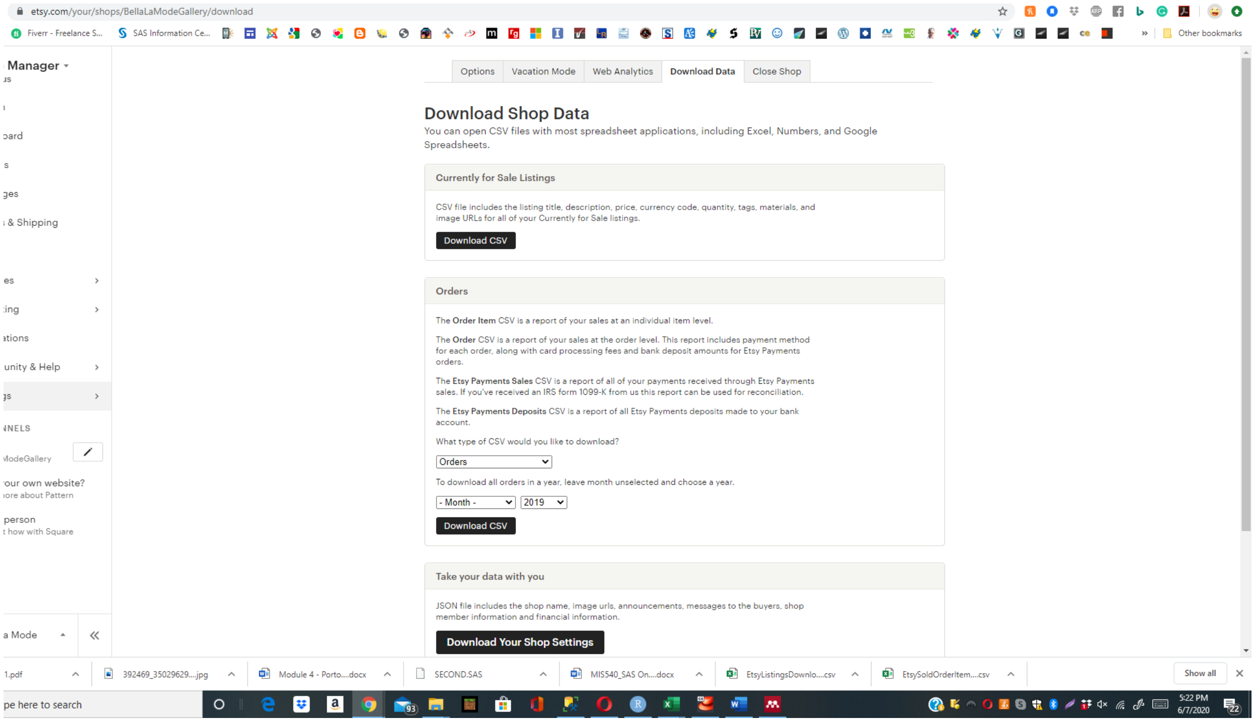
E-Commerce sales continue to grow at a rapid pace. In the second quarter of 2017 e-commerce sales of $105,096 (millions) which is 8.2% of total retail sales were reported by the United States Census Bureau (Murphy & Baer, 2017). This is a significant increase from the fourth quarter of 1999 when e-commerce sales of $5,265 (millions) or 0.7 percent of total retail sales were reported by the United States Census Bureau (Murphy & Baer, 2017). Many small businesses have gone online and run storefronts on websites like eBay and Etsy. Etsy is a global marketplace run by small businesses that sell their handcrafted and vintage goods in their online storefronts (Etsy, 2020). Etsy is in part based on a patent by M. Stinchcomb, initially filed in 2007 and granted in 2014 (Patent No. US 8,924,261 B2, 2014). One of Etsy's online storefronts called Bella La Mode Gallery has contacted an analyst to help determine how a new advertising budget can best be spent (Mode, 2020). The Analyst will examine existing data and make recommendations to the owner of Bella La Mode Gallery.

Tools Used

SAS is the tool used by the analyst on the Etsy data. SAS is top used by data analysts around the world. It originated in 1966 (SAS, 2020). The analyst will use SAS for data aggregation, data summary, data visualization, and data export. The analyst will also use a tool called Tableau to make a map.

Data Set

The data sets analyzed in this research are 2019 and 2020 orders for Bella La Mode Gallery. They were obtained from the Etsy website that the owner of Bella La Mode Gallery uses to manage the storefront.



Download Shop Data

You can open CSV files with most spreadsheet applications, including Excel, Numbers, and Google Spreadsheets.

Currently for Sale Listings

CSV file includes the listing title, description, price, currency code, quantity, tags, materials, and image URLs for all of your Currently for Sale listings.

Orders

The **Order Item** CSV is a report of your sales at an individual item level.

The **Order** CSV is a report of your sales at the order level. This report includes payment method for each order, along with card processing fees and bank deposit amounts for Etsy Payments orders.

The **Etsy Payments Sales** CSV is a report of all of your payments received through Etsy Payments sales. If you've received an IRS form 1099-K from us this report can be used for reconciliation.

The **Etsy Payments Deposits** CSV is a report of all Etsy Payments deposits made to your bank account.

What type of CSV would you like to download?

To download all orders in a year, leave month unselected and choose a year.

Take your data with you

JSON file includes the shop name, image urls, announcements, messages to the buyers, shop member information and financial information.

Figure 1: The download shop data page that allows the owner to get order data for 2019 and 2020

Data Activities

Importing the Data to SAS

The analyst is running SAS Studio as a Virtual Machine (VM) and first has to move the 2019 and 2020 order CSV files to a folder accessible by the VM. Etsy provides a separate CSV file for each year. The analyst wishes to combine the CSV data into one file for summary statistics.

```
FILENAME REFFILE '/folders/myfolders/EtsySoldOrders2019.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=Orders2019;
       GETNAMES=YES;
RUN;

PROC SORT Data=Orders2019;
       BY Order_ID;
RUN;

FILENAME REFFILE '/folders/myfolders/EtsySoldOrders2020.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=Orders2020;
       GETNAMES=YES;
RUN;
```

Figure 2: SAS Code to import the order for 2019 and Orders for 2020

The next step is to order and combine the 2019 and 2020 data into one data file.

```
PROC SORT Data=Orders2020;
       BY Order_ID;
RUN;

DATA OrdersCombined;
       MERGE Orders2019 Orders2020;
       BY Order_ID;

PROC PRINT DATA=OrdersCombined;
RUN;
```

Figure 3: SAS Code that orders and combines the 2019 and 2020 order into one file

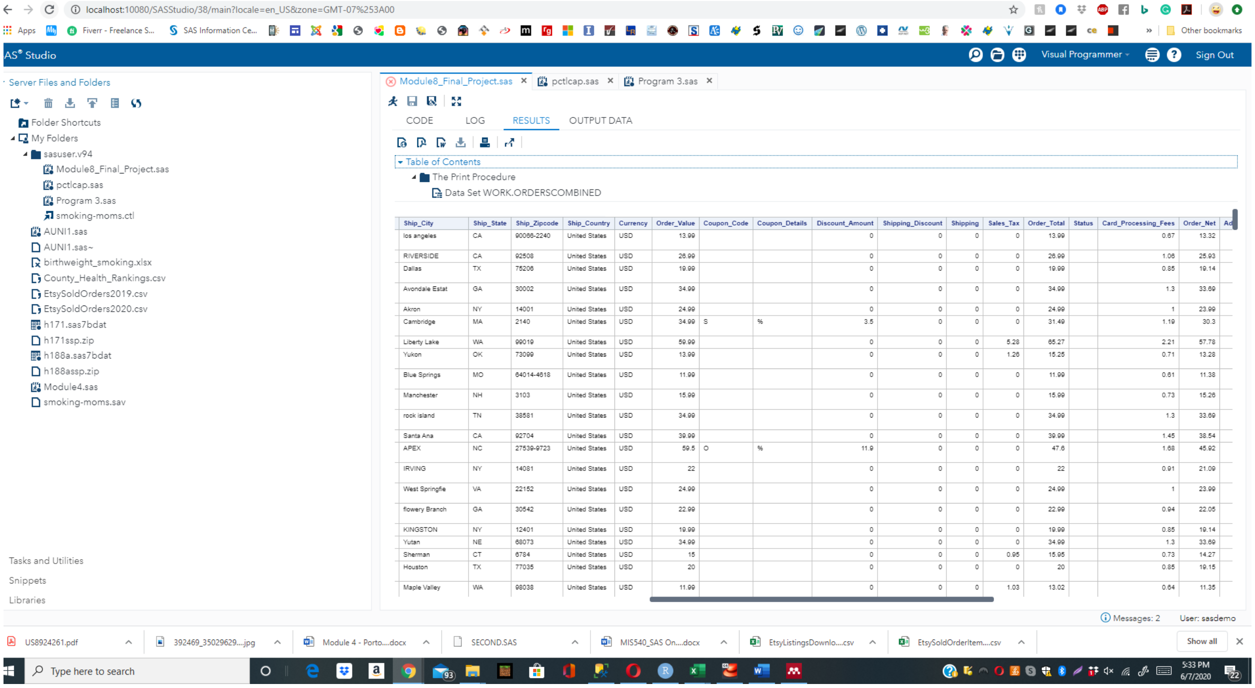


Table of Contents

The Print Procedure

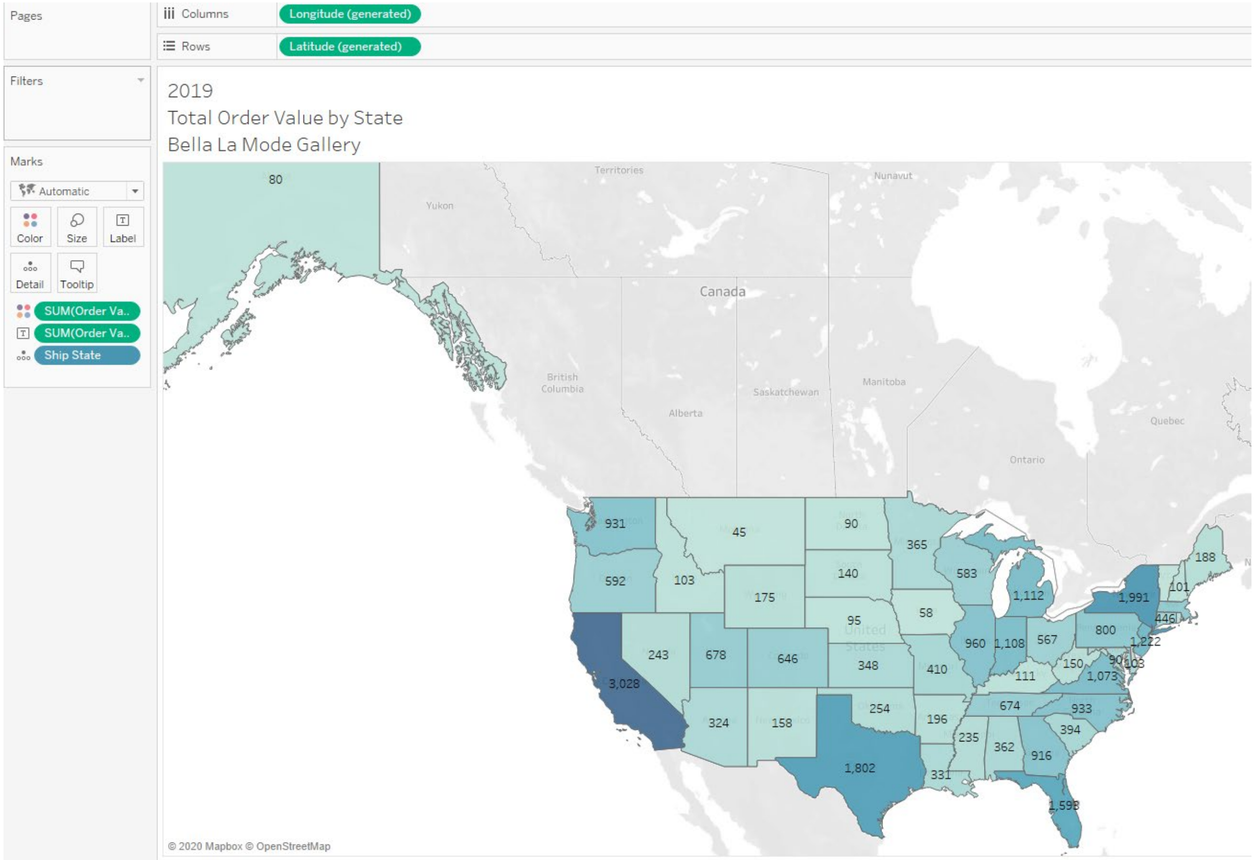
Data Set WORK.ORDERSCOMBINED

Ship_City	Ship_State	Ship_Zipcode	Ship_Country	Currency	Order_Value	Coupon_Code	Coupon_Details	Discount_Amount	Shipping_Discount	Shipping	Sales_Tax	Order_Total	Status	Card_Processing_Fees	Order_Net	Ad
los angeles	CA	90066-2240	United States	USD	13.99			0	0	0	0	13.99		0.67	13.32	
RIVERSIDE	CA	92508	United States	USD	26.99			0	0	0	0	26.99		1.06	25.93	
Dallas	TX	75206	United States	USD	19.99			0	0	0	0	19.99		0.85	19.14	
Avondale Estat	GA	30002	United States	USD	34.99			0	0	0	0	34.99		1.3	33.69	
Akron	NY	14001	United States	USD	24.99			0	0	0	0	24.99		1	23.99	
Cambridge	MA	2140	United States	USD	34.99	S	%	3.5	0	0	0	31.49		1.19	30.3	
Liberty Lake	WA	99019	United States	USD	59.99			0	0	0	5.28	65.27		2.21	57.78	
Yukon	OK	73099	United States	USD	13.99			0	0	0	1.26	15.25		0.71	13.28	
Blue Springs	MO	64014-4618	United States	USD	11.99			0	0	0	0	11.99		0.61	11.38	
Manchester	NH	3103	United States	USD	15.99			0	0	0	0	15.99		0.73	15.26	
rock island	TN	38581	United States	USD	34.99			0	0	0	0	34.99		1.3	33.69	
Santa Ana	CA	92704	United States	USD	39.99			0	0	0	0	39.99		1.45	38.54	
APEX	NC	27539-9723	United States	USD	59.5	O	%	11.9	0	0	0	47.6		1.68	45.92	
IRVING	NY	14081	United States	USD	22			0	0	0	0	22		0.91	21.09	
West Springfie	VA	22152	United States	USD	24.99			0	0	0	0	24.99		1	23.99	
flowery Branch	GA	30542	United States	USD	22.99			0	0	0	0	22.99		0.94	22.05	
KINGSTON	NY	12401	United States	USD	19.99			0	0	0	0	19.99		0.85	19.14	
Yutan	NE	68073	United States	USD	34.99			0	0	0	0	34.99		1.3	33.69	
Sherman	CT	6784	United States	USD	15			0	0	0	0.95	15.95		0.73	14.27	
Houston	TX	77035	United States	USD	20			0	0	0	0	20		0.85	19.15	
Maple Valley	WA	98038	United States	USD	11.99			0	0	0	1.03	13.02		0.64	11.35	

Figure 4: The output of the print statement on the combined orders data

The owner of Bella La Mode Gallery wants the analyst to help her make a decision. The Gallery has a limited advertising budget. The advertising campaign can be limited by geographic region or could be advertised to the whole country. The owner of the Gallery wants to know how she should spend the advertising budget.

The first thing the analyst does is a visualization in Tableau using 2019 data to determine what state has the customers who make the most purchases.



Columns: Longitude (generated)
Rows: Latitude (generated)

Marks: Automatic
SUM(Order Va..
SUM(Order Va..
Ship State

2019
Total Order Value by State
Bella La Mode Gallery

© 2020 Mapbox © OpenStreetMap

Figure 5: 2019 Total Order Value by State for Bella La Mode Gallery

The analyst notices that California is the darkest state and that the gallery did $3028 worth of sales in California for 2019. The first idea the Analyst has is that Bella La Mode Gallery should spend their entire advertising budget in California. To do this the analyst will need to split the data set into 2 groups. The first group is where the state is California and the second group is where the state is not California.

Hypothesis on Order Value

H_0 = Customers in California do not spend statistically significant more on orders than customers from the rest of the country, at Bella La Mode Galley where $\alpha = .05$

$H\alpha$ = Customers in California do spend statistically significant more on orders then customers from the rest of the country, at Bella La Mode Gallery where $\alpha = .05$

Exploring the Order Value Variable and Checking the Distribution

Order Value at Bella La Mode Gallery is the sum cost of the items ordered, not including shipping, discounts, or taxes. It is the measure of what Bella La Mode Gallery charges for an order. Before the analyst can determine what test to run to test the first hypothesis the analyst must analyze the data and see if there are outliers, and how it is distributed.

In SAS the analyst runs the proc univariate to check the normality of the data.

```
proc univariate data=OrdersCombined normal;
              VAR Order_Value;
              qqplot Order_Value /Normal(mu=est sigma=est color=red
l=1);
              run;
```

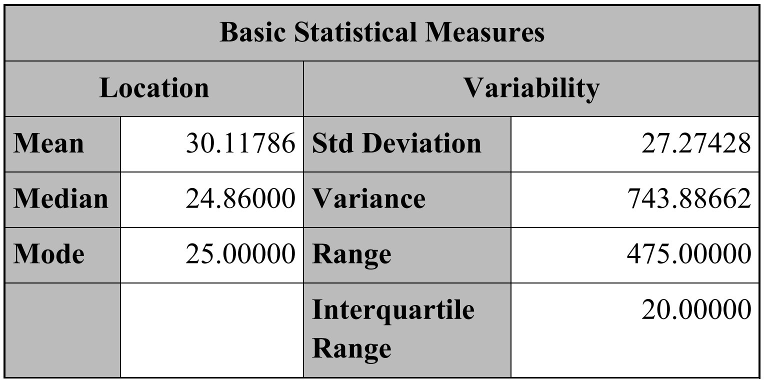
Figure 6: Running proc univariate to check the normality of the data

Selected output of Proc Univariate on variable Order_Value

Moments			
N	1302	**Sum Weights**	1302
Mean	30.1178648	**Sum Observations**	39213.46
Std Deviation	27.2742849	**Variance**	743.886616
Skewness	6.09046554	**Kurtosis**	70.8879233
Uncorrected SS	2148822.18	**Corrected SS**	967796.488
Coeff Variation	90.5584943	**Std Error Mean**	0.75587134

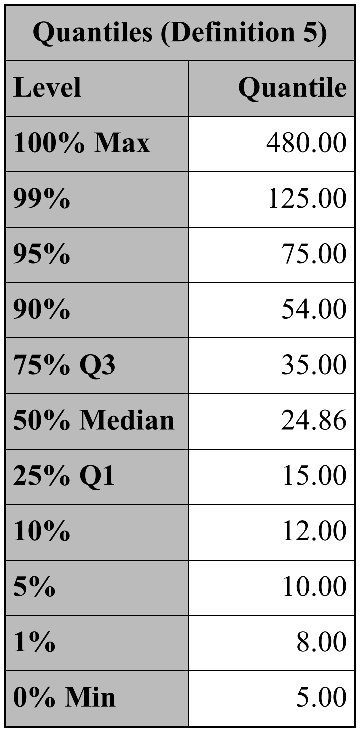
Figure 7: Moments where (n=1302, mean=30.12, Kurtosis = 70.89)

Kurtosis of 0 would mean the distribution is normal. Kurtosis of 70 means the distribution is skewed right.



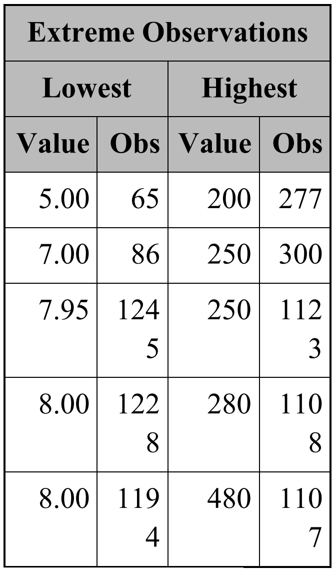
Basic Statistical Measures			
Location		Variability	
Mean	30.11786	Std Deviation	27.27428
Median	24.86000	Variance	743.88662
Mode	25.00000	Range	475.00000
		Interquartile Range	20.00000

Figure 8: Basic Statistical measures on variable order_value



Quantiles (Definition 5)	
Level	Quantile
100% Max	480.00
99%	125.00
95%	75.00
90%	54.00
75% Q3	35.00
50% Median	24.86
25% Q1	15.00
10%	12.00
5%	10.00
1%	8.00
0% Min	5.00

Figure 9: Quantiles for variable order_value



Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
5.00	65	200	277
7.00	86	250	300
7.95	1245	250	1123
8.00	1228	280	1108
8.00	1194	480	1107

Figure 10: Outliers

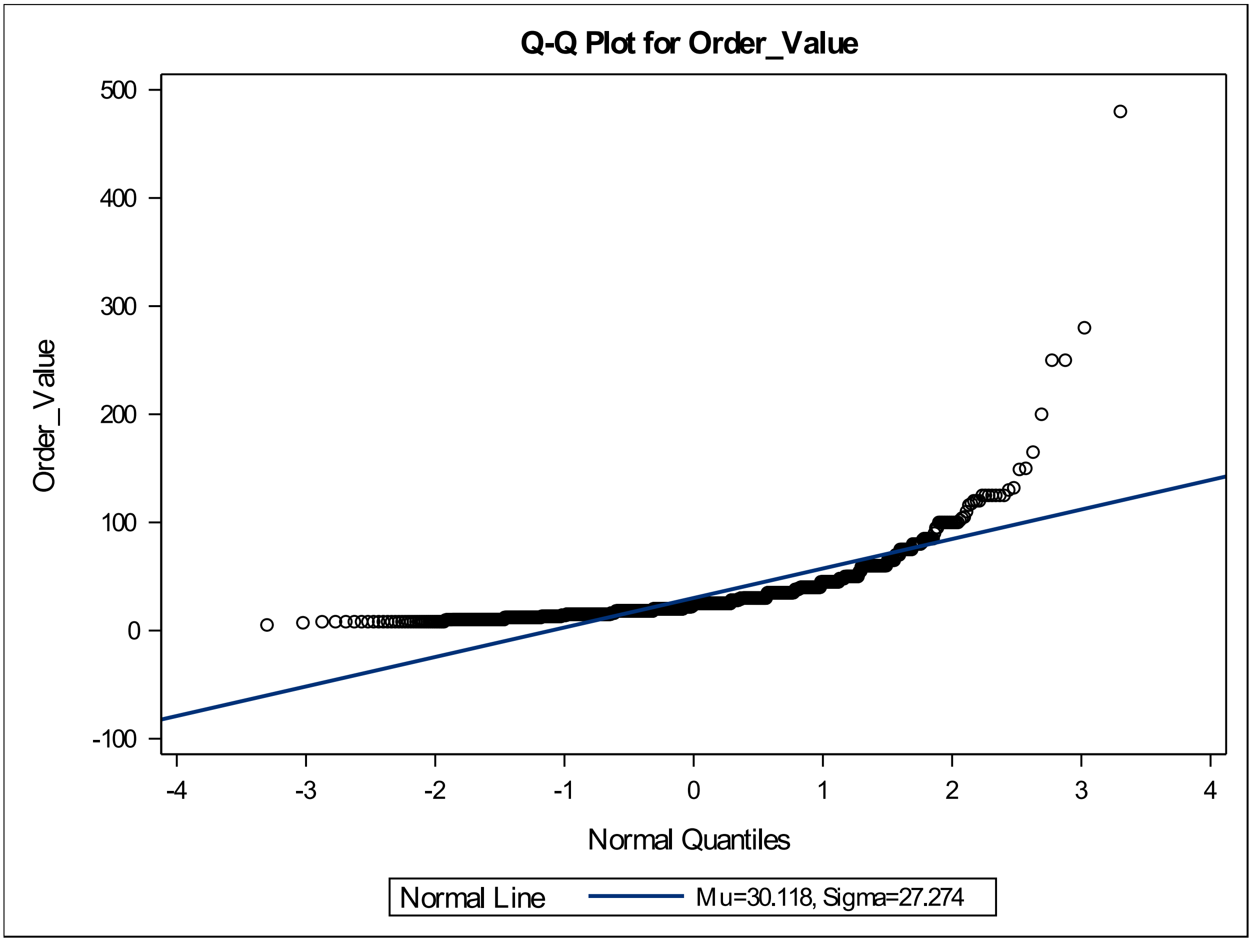
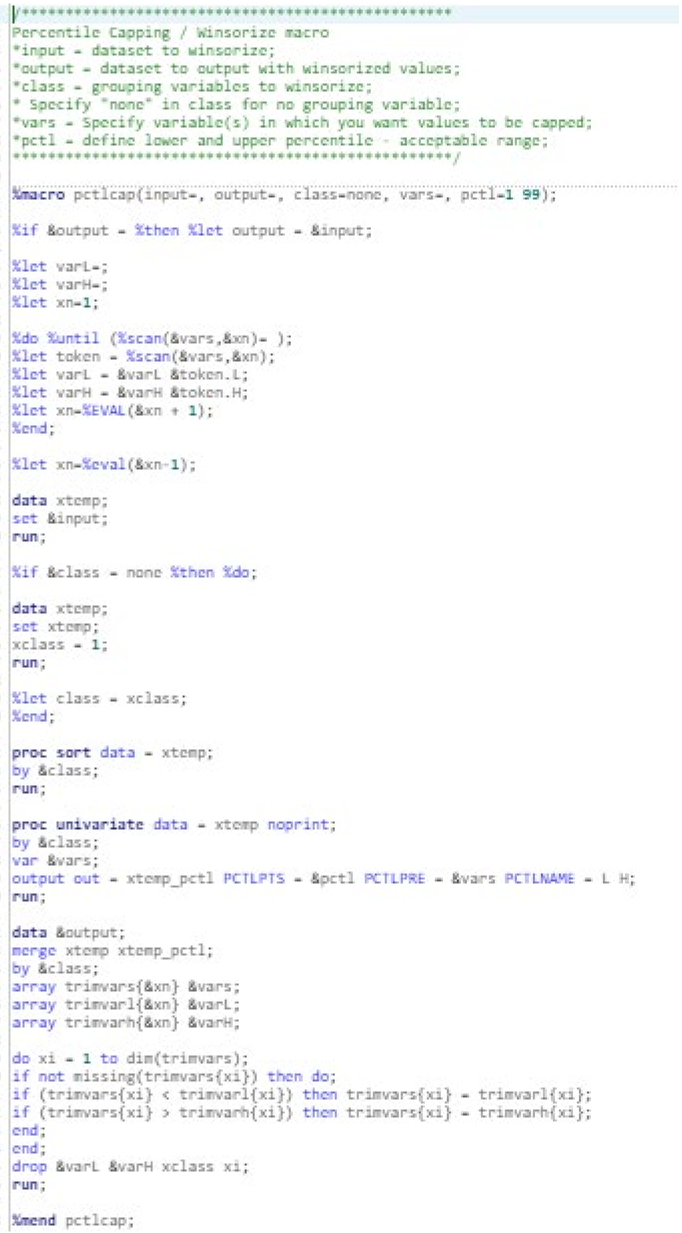


Figure 11: Q-Q Plot for Order Value

As we can see the order values do not follow a normal distribution. The outliers on each extreme are contributing to this.

Removing Outliers

After looking at the output of proc univariate the analyst decides to remove the outliers using a macro to see how close to a normal distribution can be had.



```
/****************************************************
Percentile Capping / Winsorize macro
*input = dataset to winsorize;
*output = dataset to output with winsorized values;
*class = grouping variables to winsorize;
* Specify "none" in class for no grouping variable;
*vars = Specify variable(s) in which you want values to be capped;
*pctl = define lower and upper percentile - acceptable range;
****************************************************/

%macro pctlcap(input=, output=, class=none, vars=, pctl=1 99);

%if &output = %then %let output = &input;

%let varL=;
%let varH=;
%let xn=1;

%do %until (%scan(&vars,&xn)= );
%let token = %scan(&vars,&xn);
%let varL = &varL &token.L;
%let varH = &varH &token.H;
%let xn=%EVAL(&xn + 1);
%end;

%let xn=%eval(&xn-1);

data xtemp;
set &input;
run;

%if &class = none %then %do;

data xtemp;
set xtemp;
xclass = 1;
run;

%let class = xclass;
%end;

proc sort data = xtemp;
by &class;
run;

proc univariate data = xtemp noprint;
by &class;
var &vars;
output out = xtemp_pctl PCTLPTS = &pctl PCTLPRE = &vars PCTLNAME = L H;
run;

data &output;
merge xtemp xtemp_pctl;
by &class;
array trimvars{&xn} &vars;
array trimvarl{&xn} &varL;
array trimvarh{&xn} &varH;

do xi = 1 to dim(trimvars);
if not missing(trimvars{xi}) then do;
if (trimvars{xi} < trimvarl{xi}) then trimvars{xi} = trimvarl{xi};
if (trimvars{xi} > trimvarh{xi}) then trimvars{xi} = trimvarh{xi};
end;
end;
drop &varL &varH xclass xi;
run;

%mend pctlcap;
```

Figure 12: Macro for Percentile Capping / Winsorizing (Bhalla, 2020)

The macro shown in figure 10 will accomplish winsorization of the data. Winsorization is when any values above the 99th quantile are replaced with the value in the 99th quantile and any values below the 1st quantile are replaces with the value in the 1st quantile (Bhalla, 2020). Figure 9 shows us the 100% max quantile is 425 so we will have a bigger effect on the upper end of the distribution versus the 0% mine quantile being replaced by the 1% quantile.

```
*remove outliers below the 1% quantile and 99% quantile - winsorization;

%pctlcap(input=OrdersCombined, output=OrdersNoOutliers, class=none, vars = Order_Value,

pctl=1 99);
```

Figure 13: Running the macro to do the winsorization

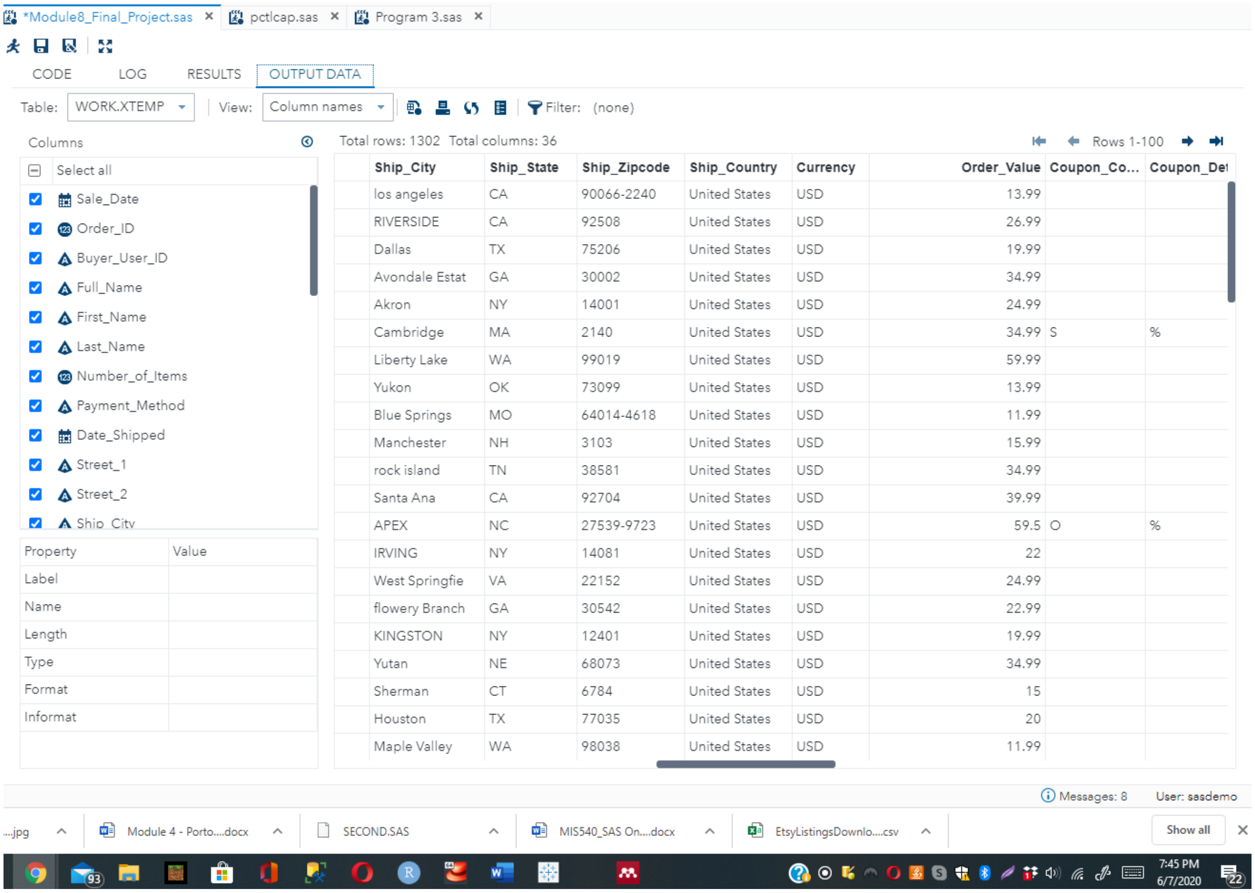


Table: WORK.XTEMP View: Column names Filter: (none)

Columns

- Select all
- ☑ Sale_Date
- ☑ Order_ID
- ☑ Buyer_User_ID
- ☑ Full_Name
- ☑ First_Name
- ☑ Last_Name
- ☑ Number_of_Items
- ☑ Payment_Method
- ☑ Date_Shipped
- ☑ Street_1
- ☑ Street_2
- ☑ Ship_City

Property	Value
Label	
Name	
Length	
Type	
Format	
Informat	

Total rows: 1302 Total columns: 36 Rows 1-100

Ship_City	Ship_State	Ship_Zipcode	Ship_Country	Currency	Order_Value	Coupon_Co...	Coupon_Det
los angeles	CA	90066-2240	United States	USD	13.99		
RIVERSIDE	CA	92508	United States	USD	26.99		
Dallas	TX	75206	United States	USD	19.99		
Avondale Estat	GA	30002	United States	USD	34.99		
Akron	NY	14001	United States	USD	24.99		
Cambridge	MA	2140	United States	USD	34.99	S	%
Liberty Lake	WA	99019	United States	USD	59.99		
Yukon	OK	73099	United States	USD	13.99		
Blue Springs	MO	64014-4618	United States	USD	11.99		
Manchester	NH	3103	United States	USD	15.99		
rock island	TN	38581	United States	USD	34.99		
Santa Ana	CA	92704	United States	USD	39.99		
APEX	NC	27539-9723	United States	USD	59.5	O	%
IRVING	NY	14081	United States	USD	22		
West Springfie	VA	22152	United States	USD	24.99		
flowery Branch	GA	30542	United States	USD	22.99		
KINGSTON	NY	12401	United States	USD	19.99		
Yutan	NE	68073	United States	USD	34.99		
Sherman	CT	6784	United States	USD	15		
Houston	TX	77035	United States	USD	20		
Maple Valley	WA	98038	United States	USD	11.99		

Figure 14: Output order data after it has had outliers removed (winsorization)

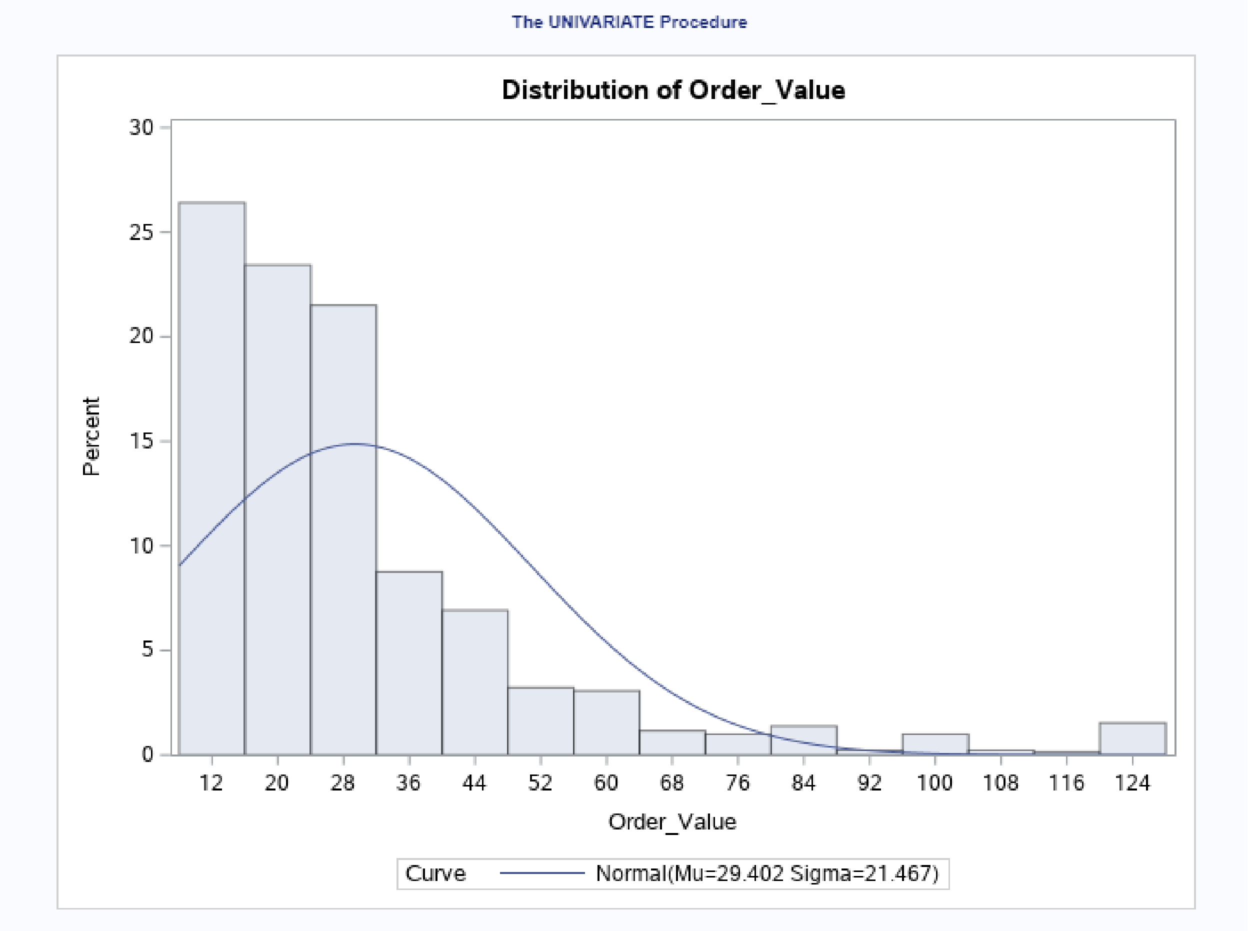


Figure 15: Histogram showing the distribution of order value after removing outliers

Next, the analyst will split the data into two categories.

```
DATA OrdersByState;

          SET OrdersNoOutliers;


          StateCategory = "";

          If Ship_State="CA" THEN StateCategory = "C"; ELSE StateCategory = "A";


RUN;
```

Figure 16: SAS code to add a new field to categorize the states

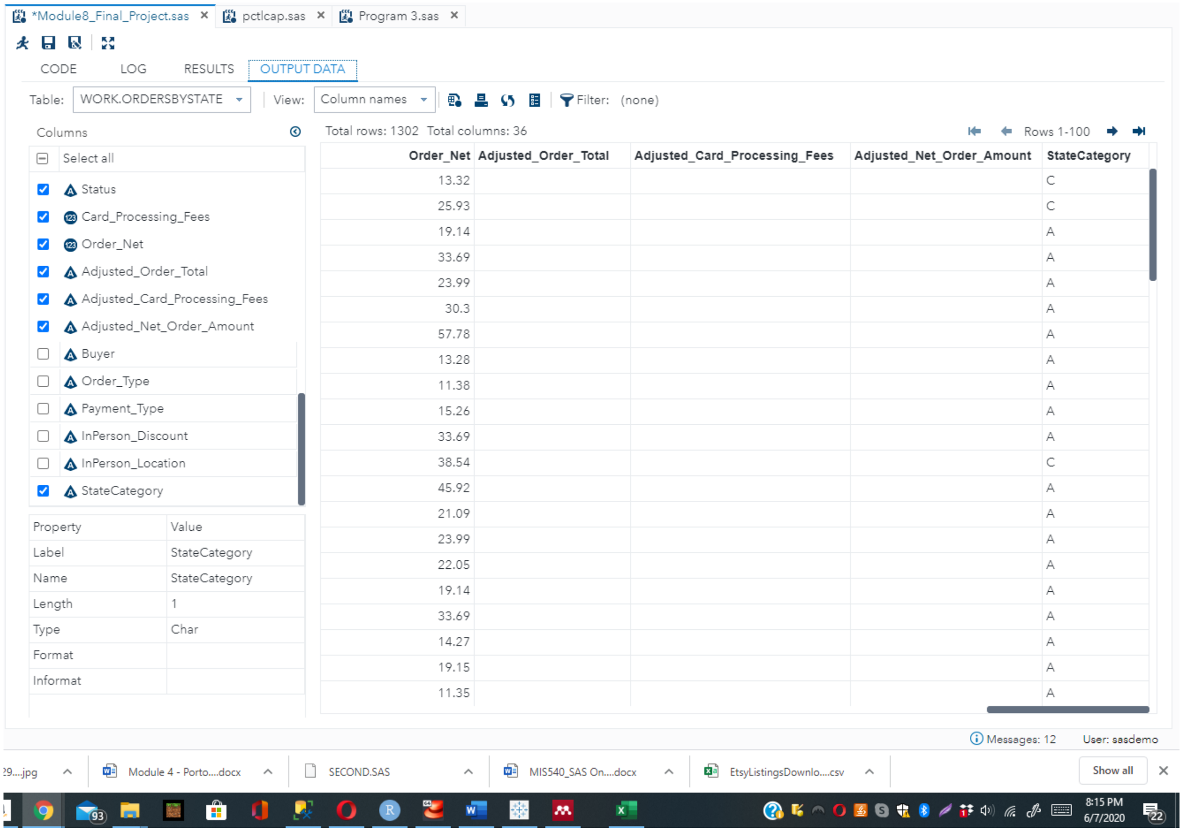


Table: WORK.ORDERSBYSTATE View: Column names Filter: (none)

Columns

- ☐ Select all
- ☑ Status
- ☑ Card_Processing_Fees
- ☑ Order_Net
- ☑ Adjusted_Order_Total
- ☑ Adjusted_Card_Processing_Fees
- ☑ Adjusted_Net_Order_Amount
- ☐ Buyer
- ☐ Order_Type
- ☐ Payment_Type
- ☐ InPerson_Discount
- ☐ InPerson_Location
- ☑ StateCategory

Property	Value
Label	StateCategory
Name	StateCategory
Length	1
Type	Char
Format	
Informat	

Total rows: 1302 Total columns: 36 Rows 1-100

Order_Net	Adjusted_Order_Total	Adjusted_Card_Processing_Fees	Adjusted_Net_Order_Amount	StateCategory
13.32				C
25.93				C
19.14				A
33.69				A
23.99				A
30.3				A
57.78				A
13.28				A
11.38				A
15.26				A
33.69				A
38.54				C
45.92				A
21.09				A
23.99				A
22.05				A
19.14				A
33.69				A
14.27				A
19.15				A
11.35				A

Messages: 12 User: sasdemo

29....jpg Module 4 - Porto....docx SECOND.SAS MIS540_SAS On....docx EtsyListingsDownlo....csv

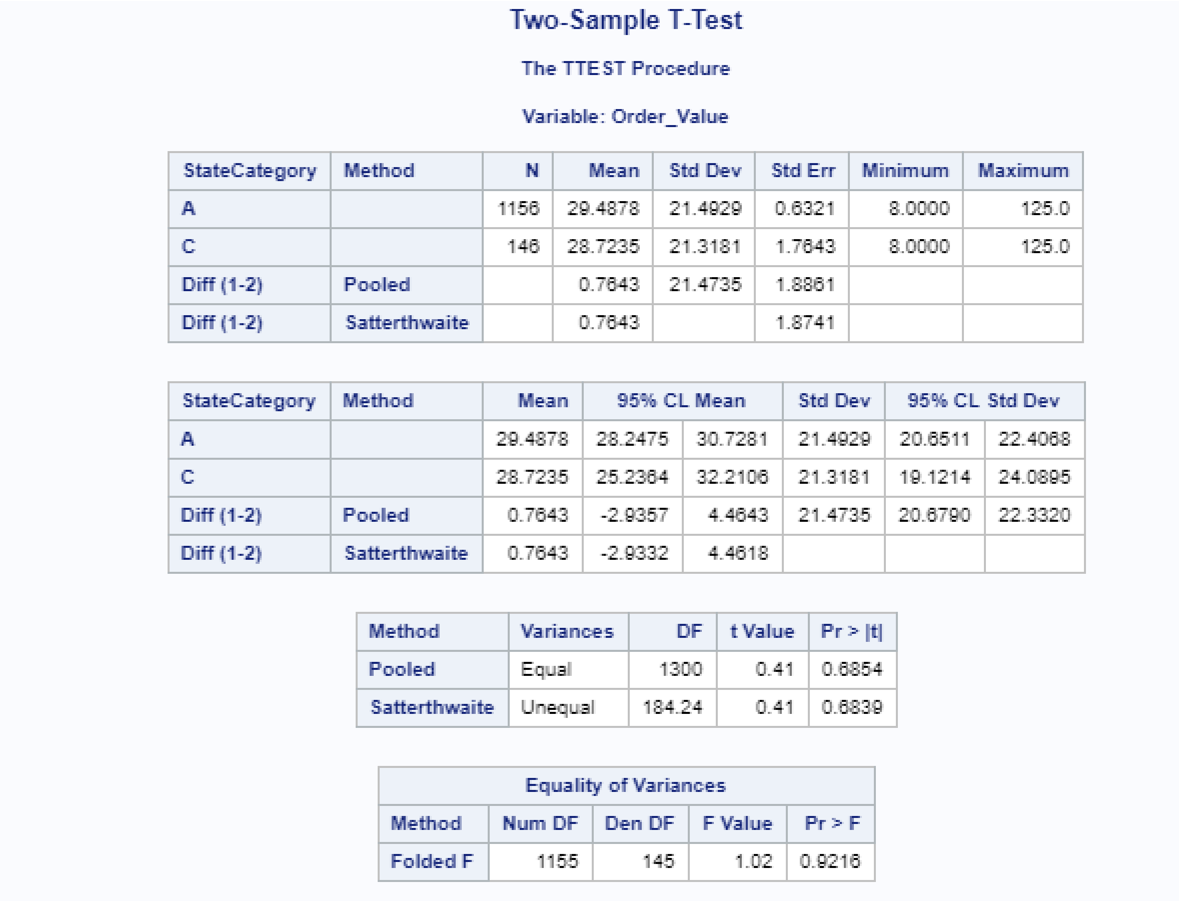
8:15 PM 6/7/2020

Figure 17: New data set called OrdersNoOutliers that is categorized by C for California or A for all others.

The next step is to run a two-sample t-test to test the Hypothesis on Order Value presented earlier.

```
* Run a 2 sample test based on StateCategory and Order_Value;

PROC TTEST;

CLASS StateCategory;

VAR Order_Value;

Title "Two-Sample T-Test";

RUN;
```

Figure 18: Code to run the two-sample t-test



Two-Sample T-Test

The TTEST Procedure

Variable: Order_Value

StateCategory	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
A		1156	29.4878	21.4929	0.6321	8.0000	125.0
C		146	28.7235	21.3181	1.7643	8.0000	125.0
Diff (1-2)	Pooled		0.7643	21.4735	1.8861		
Diff (1-2)	Satterthwaite		0.7643		1.8741		

StateCategory	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
A		29.4878	28.2475	30.7281	21.4929	20.6511	22.4068
C		28.7235	25.2364	32.2106	21.3181	19.1214	24.0895
Diff (1-2)	Pooled	0.7643	-2.9357	4.4643	21.4735	20.6790	22.3320
Diff (1-2)	Satterthwaite	0.7643	-2.9332	4.4618			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1300	0.41	0.6854
Satterthwaite	Unequal	184.24	0.41	0.6839

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1155	145	1.02	0.9216

\

Figure 19: Output from the two-sample t-test

The two-sample t-test output shows strong evidence that the means are similar enough between state category c (California) and state category A(all others),

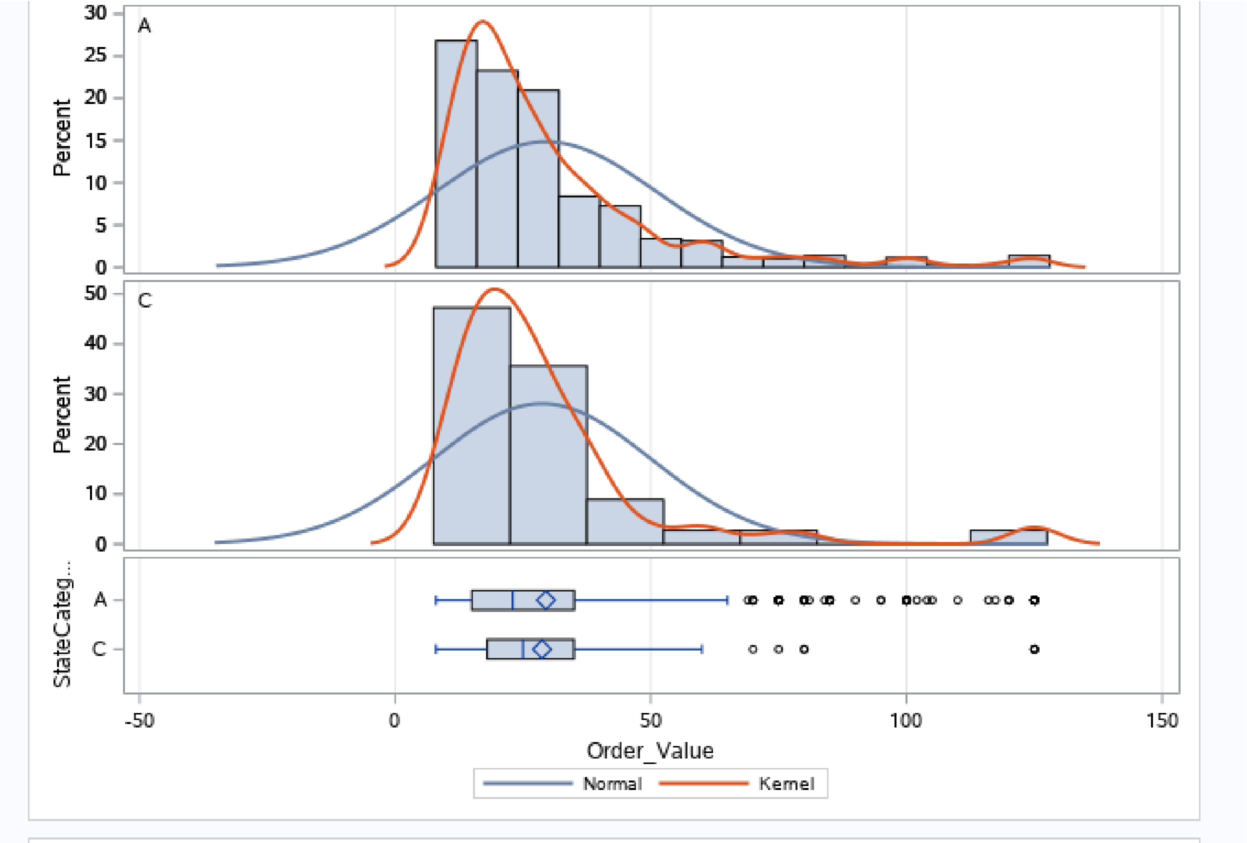


Figure 20: Graphic output from the two-sample t-test

The graphic in figure 20 shows that the distributions of order value are similar from orders placed in the state of California to the orders placed in the remainder of the united states.

Analytical Outcome

Based on the evidence gained by viewing the results of the two-sample t-test we do not have enough evidence to reject the null hypothesis. Thus, we affirm the null hypothesis of H_0 = Customers in California do not spend statistically significant more on orders than customers from the rest of the country, at Bella La Mode Galley where α = .05. Considering this the analyst will let the owner of Bella La Mode Gallery know that the advertising campaign can target customers from all of the United States. There is no reason to just advertise to California.

Conclusion

Bella La Mode Gallery is a small e-commerce store that runs on the Etsy platform. E-Commerce continues to grow as a percentage of retail sales year after year (Murphy & Baer, 2017). The analyst in this case though it might make sense to spend a limited advertising budget just advertising to California since the Tableau map showed the majority of sales originating in California. However, further statistical analysis revealed no significant difference in the mean order value between orders placed in California and orders placed in all other states. Therefore the owner of Bella La Mode Gallery was advised to not use the geographic state of order original as a variable to influence where marketing campaigns occur.

References

Bhalla, D. (2020). DETECTING AND SOLVING PROBLEM OF OUTLIER. Retrieved from Listen Data website: https://www.listendata.com/2015/01/detecting-and-solving-problem-of-outlier.html

Etsy. (2020). Keep Commerce Human. Retrieved July 6, 2020, from https://www.etsy.com/about?ref=ftr

Mode, B. La. (2020). Bella La Mode Gallery. Retrieved from http://www.bellalamodegallery.com

Murphy, J., & Baer, A. B. (2017). Overview of E-Commerce Statistics United States Census Bureau. *32nd Meeting of the Voorburg Group on Service Statistics*. Retrieved from http://164.100.163.194/images/Newpaperslatest20oct/US_E-Commerce_Paper.pdf

SAS. (2020). About SAS. Retrieved September 5, 2020, from https://www.sas.com/en_us/company-information/profile.html

Stinchcomb, M. (2014). *Patent No. US 8,924,261 B2*. Retrieved from https://patentimages.storage.googleapis.com/17/f0/be/48a80586754d5b/US8924261.pdf