# DAT 220 Module 8 – Final Project

10-24-2021 v1

Revised

3-26-2023

Erik Wilhelm

erik.wilhelm@snhu.edu

Southern New Hampshire University

## Abstract

This document contains my interpretation of the requirements of the assignment as well as the needs of Bubba Gump’s analyses team. I will propose my idea as to the overall business problem as well as propose the analytical methods and informational yields of employing the methods.

As of 3-26-2023 I have evaluated the dataset provided in the original report and intend to provide support and additional information using Orange Data Mining software as an enhancement to my original report. All updates and additions will be in this color of blue.

## Plan for Analysis

Background and introduction

The business problem that I see from the problem statement is that the Bubba Gump organization has seen declining sales since the movie release and “newness” of the restaurant has reduced. The problems that I plan to review and solve are as follows. After initial inspection of the responses from the test group of Bubba Gump’s customers and understanding the requirements of Bubba Gump management I feel that the overall business problems that I want to solve based on the availability of data are: Which customers are most likely to make online purchases and how to increase web traffic.

Tools and Visualizations

I plan to use JMP software to compare and display the data using various histograms, tables and charts as created from the software. JMP is a very highly regarded analytic software that eliminates the need for coding knowledge and provides an extensive tested and verified suite of visual and forecasting software. JMP software has a large following and great customer support, but it does lack a bit in the finer points of clarifying data, adding extra steps to get to a final product. The data that has been compiled from organization survey sampling. I plan to review numbers of customers from states and zip codes for in house dining/purchases as well as online sales. I plan to review web visits and web purchases as well as multiple visits (return visits) as well as age of customers. I feel that by looking at comparison charts I can view and plot the locations that are most effective at sales and repeat business. By comparing the data of online sales and in store sales we will be able to ascertain where the most sales are occurring and provide actual data on what we are doing good and where we could do better. I would review and compare data in which states the restaurants are performing well and where the online sales are being generated. Using this data, the information can be used to build business where it is currently lacking as well as show where the organization is doing well at reaching customers and bringing them to the online or brick and mortar locations.

I will be reviewing the aforementioned bubbagump dataset and will be evaluating it using Orange Data Mining version 3.34.0. Orange is n open source software that is provided for educational, research and professional use by the Bioinformatics Lab at the University of Ljubljana, Slovenia. It is released for public use and provides for community-based improvements.

Specific Research Question

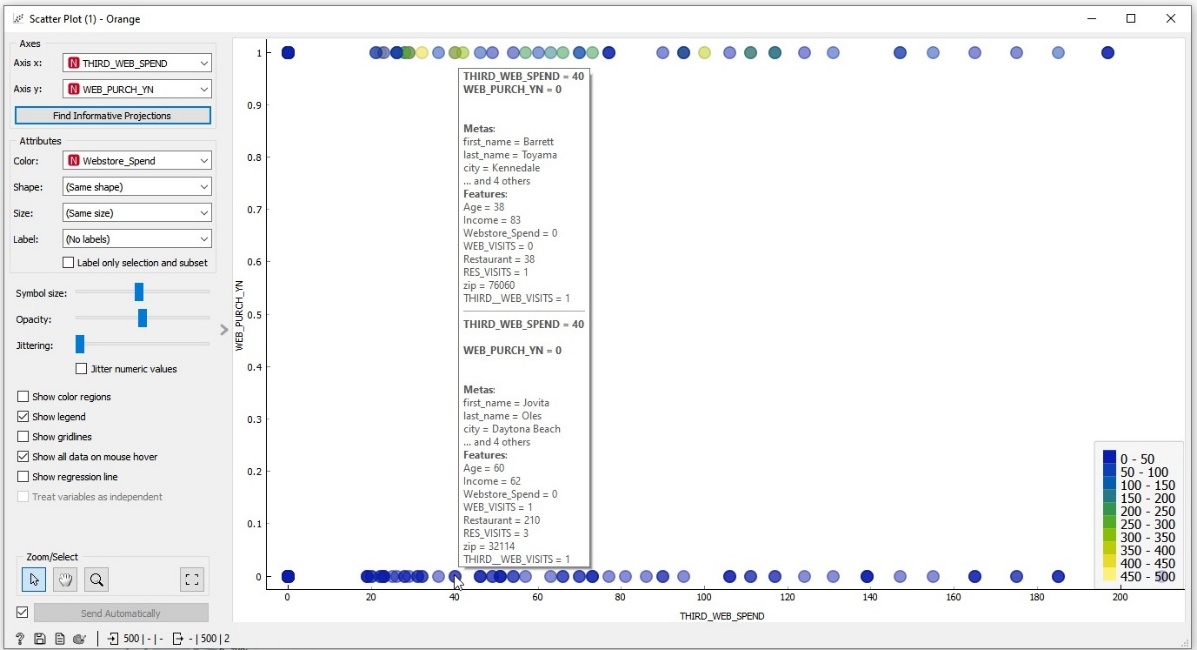
I feel that the specific business question that is to be addressed through the data analysis is~~, why sales has declined and where specifically are the declining sales stemming from~~ how to increase online sales? The question that I intend to answer with my report is, where are most sales occurring for the organization is it online or in store and geographically where are the sales the highest or lowest? By evaluating the provided data using the JMP software, I will be able to show where the most sales are being generated and which states are performing best. If I can compile data to provide significant results that point to a specific location online or in-store and where the store is located, I will be able to consider the question answered.

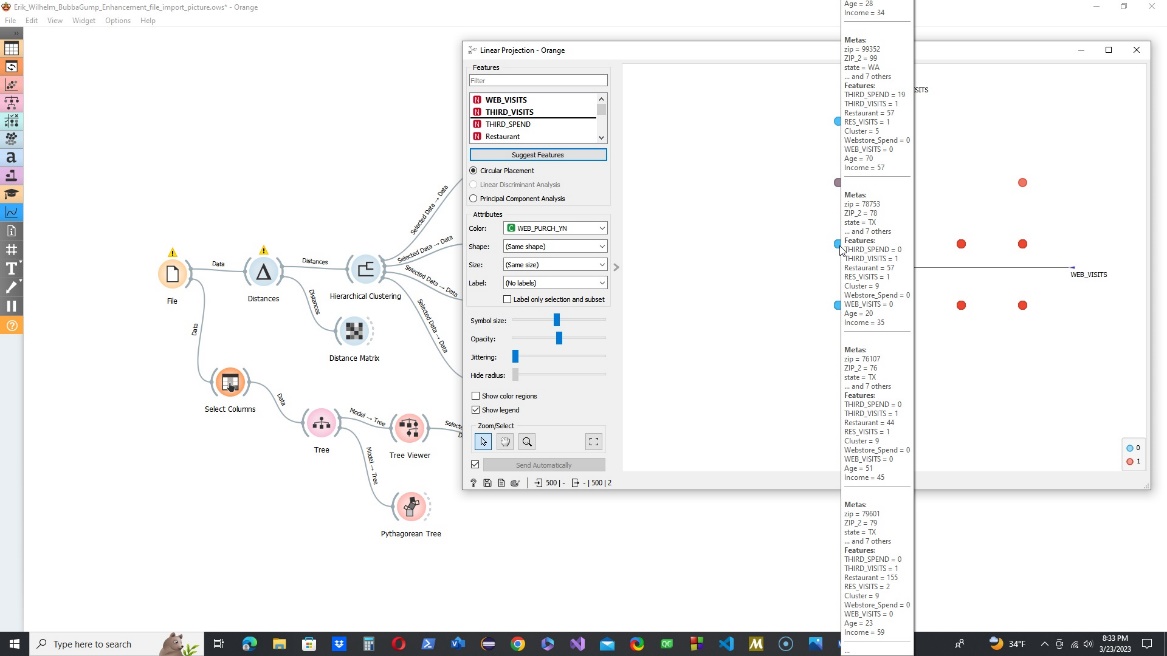
Analysis

The report that I have provided has been revised and reviewed by myself and my peers. The report has been prepared by looking at the individual variables and possible analytical concerns that each might pose to the final report. The diagrams have been built based on previous outcomes and revised from data and studies. The final report is my interpretation of the business problem at hand and the visual aids that support my findings as well as point out inconsistencies in the data to provide information for future data collection and recommendations for the Bubba Gump organization.

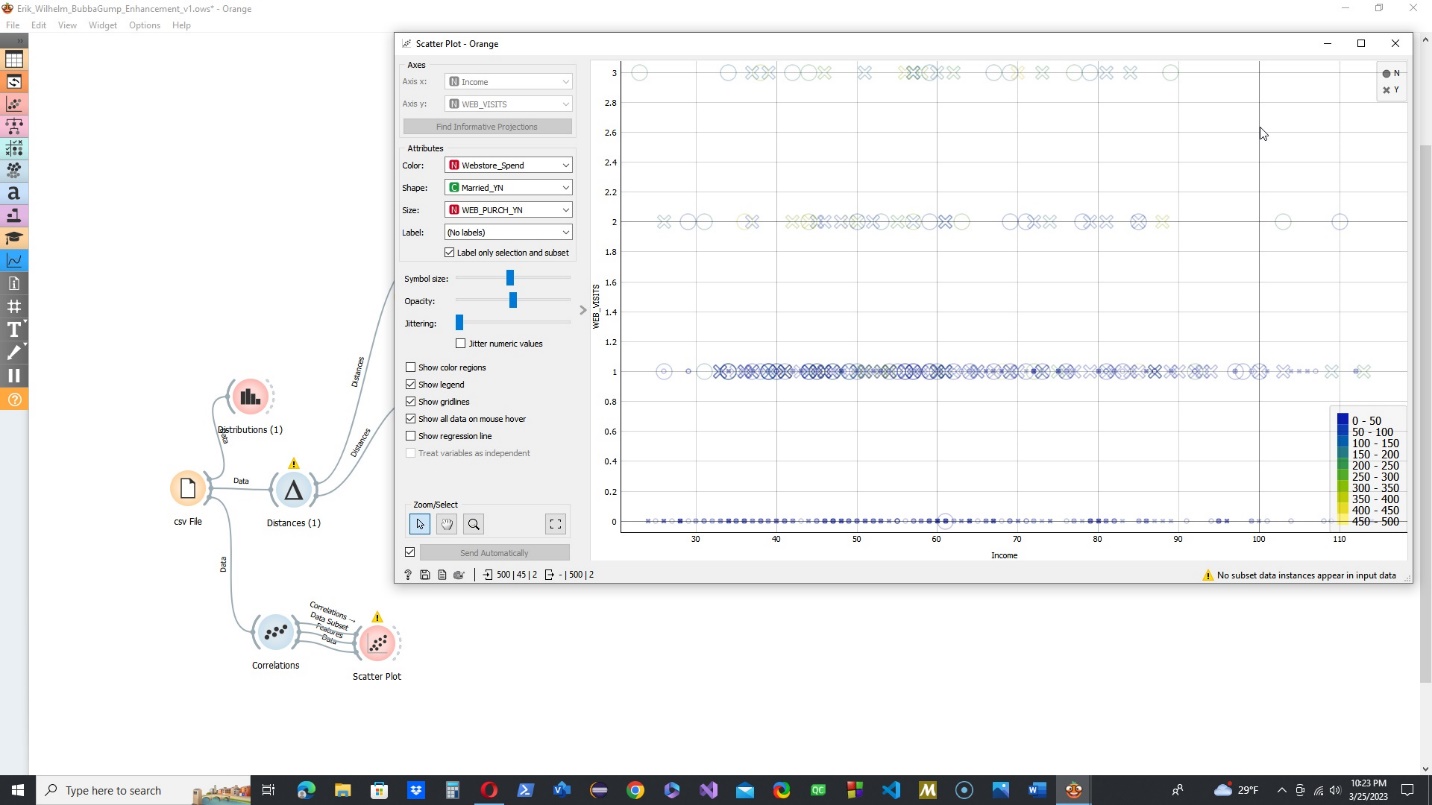
The sources of error in the data that was provided with respect to the clustering diagrams was the multiple answers to similar questions, such as zip code and zip code 2 as well as marital status yes/no and marital status binary. Another source of error comes in with how the data is provided and interpreted. I found that the zip code data and the restaurant data is provided as a number variable which is accurate, but the software interprets them as ordinal variables which causes the data to be plotted as simple numbers and losing the specificity of each number such as zip code and geographical location. ~~The main source of error comes in with how the data is provided and interpreted. I found that the zip code data and the restaurant data is provided as a number variable which is accurate, but the software interprets them as ordinal variables which causes the data to be plotted as simple numbers and losing the specificity of each number such as zip code and geographical location~~  removed redundant statement. I have used the data samples that appeared applicable to me. This allows for my own evaluation without bias but allows for my impressions onto the data~~. The main source of error comes the same as with linear regression modeling with less impact. I found that the zip code data and the restaurant data is provided as a number variable which is accurate, but the software interprets them as ordinal variables which causes the data to be plotted as simple numbers and losing the specificity of each number such as zip code and geographical location. I have used the data samples that appeared applicable to me. This allows for my own evaluation without bias but allows for my impressions onto the data.~~ Removed redundant statement. This has a bit less impact because you can see the broad view of the various data points of a less concise (nonlinear dependent) diagram. When evaluating the data using Orange I found that there were several inaccuracies in the data collection that would cause errors in possible reports. This can most notable be seen in the image below where I was able to visualize inconsistencies in web purchase vs online sales. The images below show discrepancies between online spending and online visits. The dataset was the dependent study in an effort to increase online sales but inaccurately depicts customer actions.

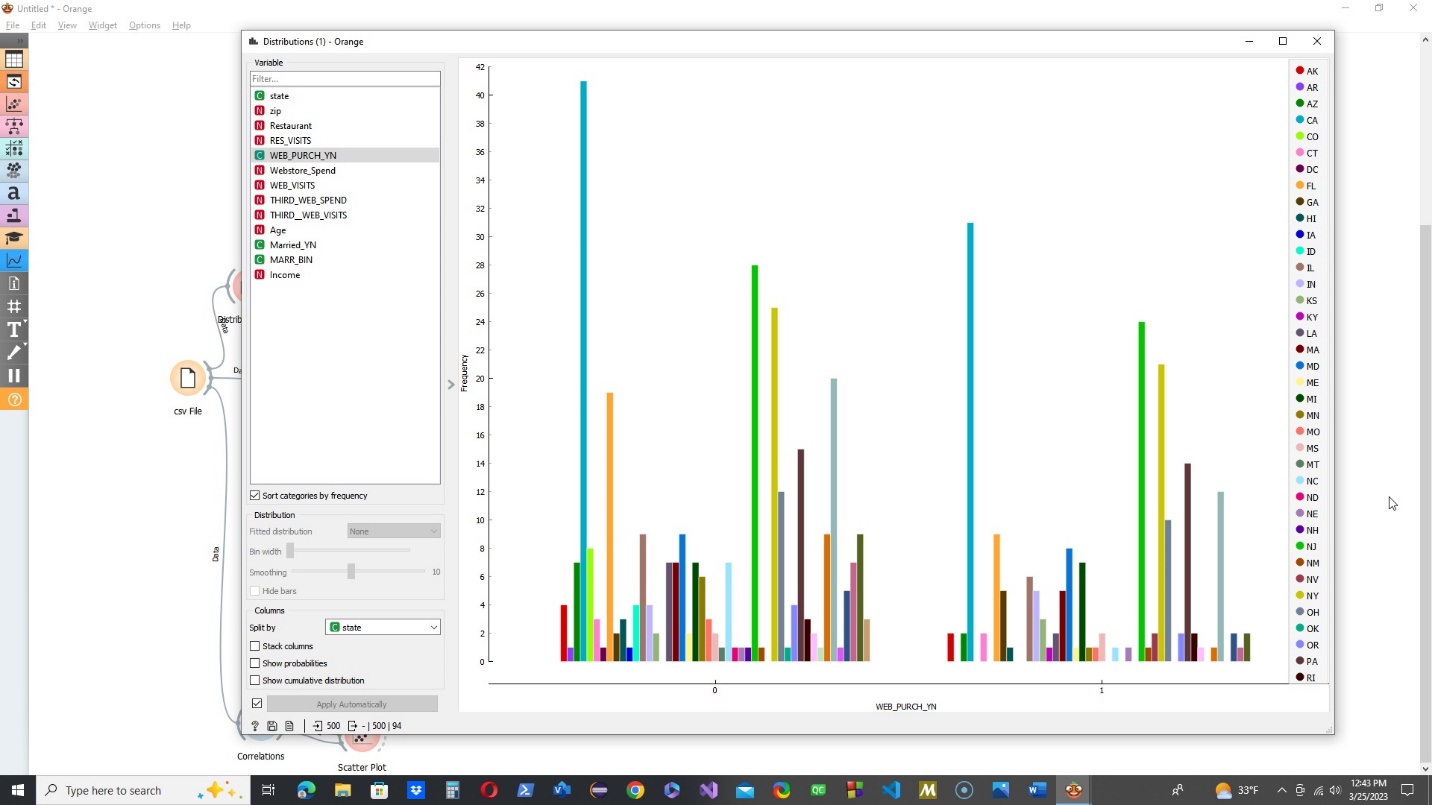
Areas of inaccuracy using Orange Data mining can be from user error as I am new to the software and have worked to understand the workings and outputs of the software but there is a learning curve in all new software. The calculations are performed under the hood so manipulation is much more difficult limiting to the widgets on hand. I have found a large support community for the software which provides for community support and functionality enhancement this is good and bad as it allows for more development opportunities but limits updates to need and allows for possible introduction of bugs. The software is built using python code so it is definitely designed with the data scientist in mind. I have not looked into creating additional options for widgets but there are numerous online areas that discuss these topics including links on the Orange website provided in the reference area.

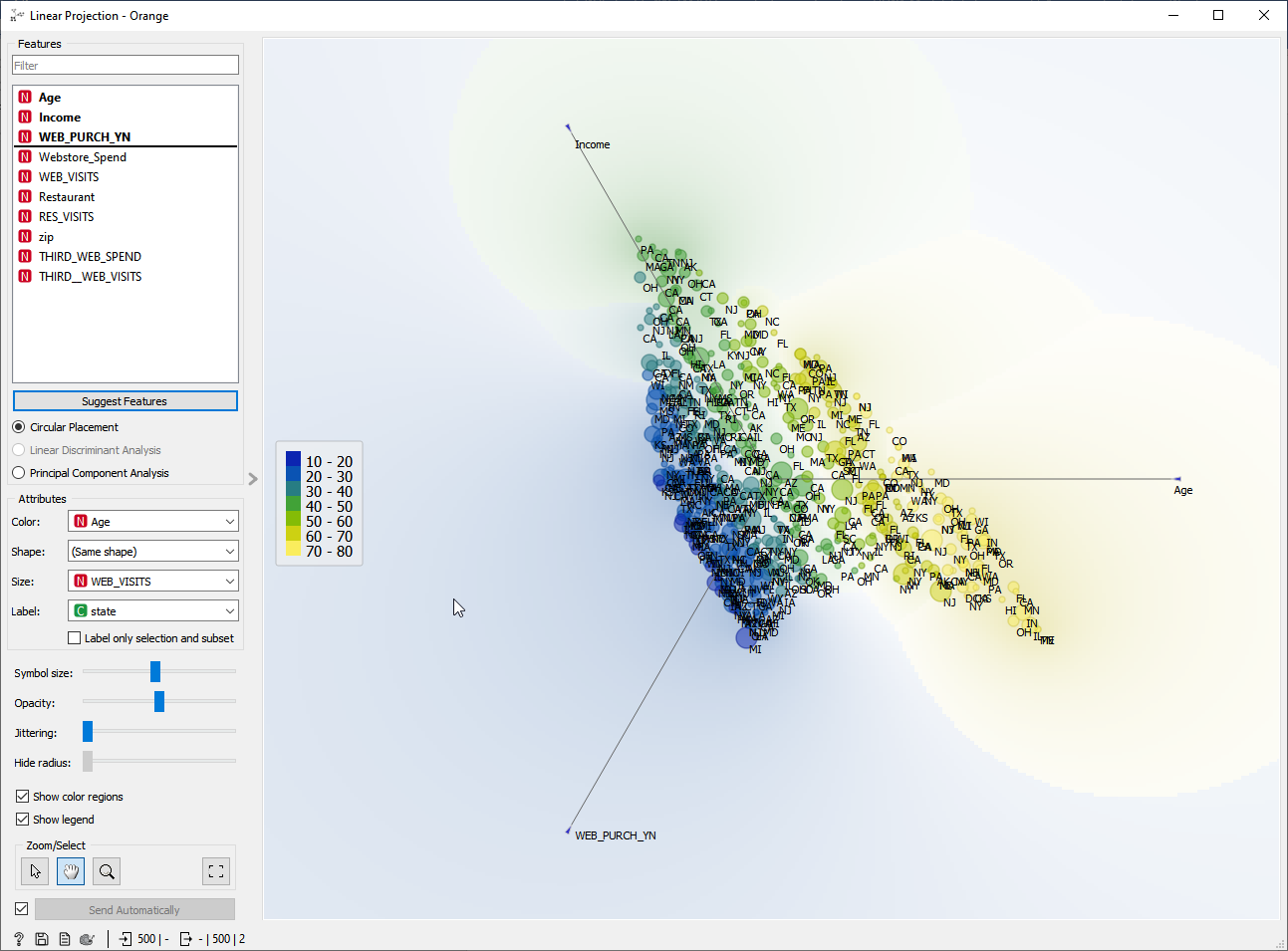


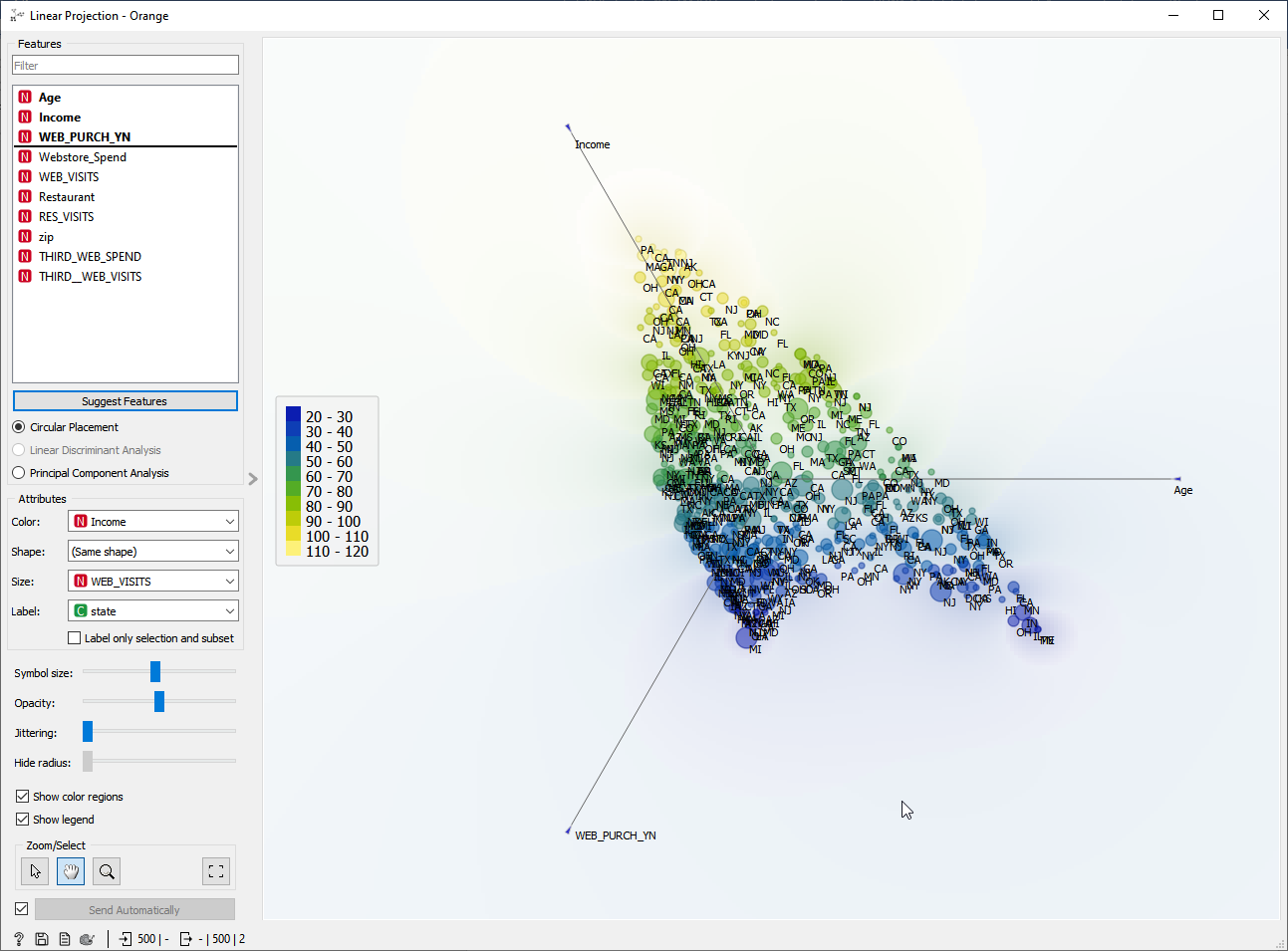


The most meaningful pattern to support the main question of getting a larger online sales presence is that the clusters significantly point to clustering that represents webstore visits, and spending. This tells the data analyst that there is reason to believe that the online store has room to grow, and the next question is how to bring people to the site as well as which people to target for increasing sales online. From the patterns that were apparent in the clustering diagrams I was able find significant pattens in web spending as a Y determinant variable and age, State, marital status, and household income. These evaluations led to an interpretation that the most likely candidates have a household income of 40-60 thousand dollars, 20-40 years old and living in New York, New Jersey, and California. These patterns indicate that the data should be further evaluated to see if it is people that live in these states or are on vacation and who the organization should target for advertising. The significant patterns that I found is that the best place to look for increasing sales is for married and unmarried patrons, between 20 and 40 years old with an annual income between 20 and 40 thousand dollars and more importantly first web visitors. This suggests that the site is not encouraging repeat visits or sales. This would be a future evaluation of how to make the web store more popular. Given the known and stated possible inaccuracies using the Orange software I have found the same trends in the data and would confirm that the prime area of focus for online sales is customers 20-40 years of age with a household income of $40-$60 thousand per year as stated in my original report.









Inaccurate depictions in the data appear like sources of error. This is most common in areas where data numbers are meant to be specific and require additional evolvement of the data. This is most prominently seen in the zip code responses where a zip code is seen by the software solely as a number and the geographic aspect is lost. This phenomenon is additionally manifested in the metrics for restaurant and marital status. To address the items, I looked for alternate data that was provided to evaluate that could provide results. Looking at Zip code I used the state provided instead and for marital I looked at data that is either 0 or 1 which plots more as 0-1 for no and 1-2 for yes. The data that would still need additional cleaning would be restaurant number and where that restaurant is. I tried to find a store number listing online but have not been successful as of now. The inaccurate depictions of data are most prominent in the linear regression modeling and required that I left several variables or data columns out of the evaluation. This is most common in answers that were provided as numbers but were meant to represent other data such as restaurant and zip code. The marital status binary and web purchase yes/no does not provide specificity, but the evaluations can pe performed using a metric of 0 to -1 for no and 1+ to 2. The inaccuracy of data depiction is that there is only one ordinal variable to evaluate logistic regression against. This could suggest to outside parties that the only important metric is web purchases. While this is a significant part of the data mining requirement the organization has specifically asked about increasing sales in store as well as online.

Based upon the results of the clustering evaluation I find that alternative methods that I will evaluate are jmp software variable clustering option as well as jmp explore patterns option. These additional evaluations will potentially lead to more insight that can aid in answering the main question of increasing web traffic and declining sales overall. The data that I chose to leave out of the linear regression is the greatest weakness of my report and I plan to look for additional methods of modifying the data or interpreting the evaluations to provide usable histograms. I would also evaluate the data by alternating the dependent and independent variables to find a pattern of stronger comparisons. The most significant alternative methods would be to include additional ordinal variables that would provide for additional logistic regression modeling and evaluation. I would also evaluate more variables against the one variable that is provided in the sample. I am not sure if variable types could be changed (or even should) but that would be an alternative option as well.

Final Report

## Display and Interpretation

## **Hierarchical Cluster (using all columns of data survey)**

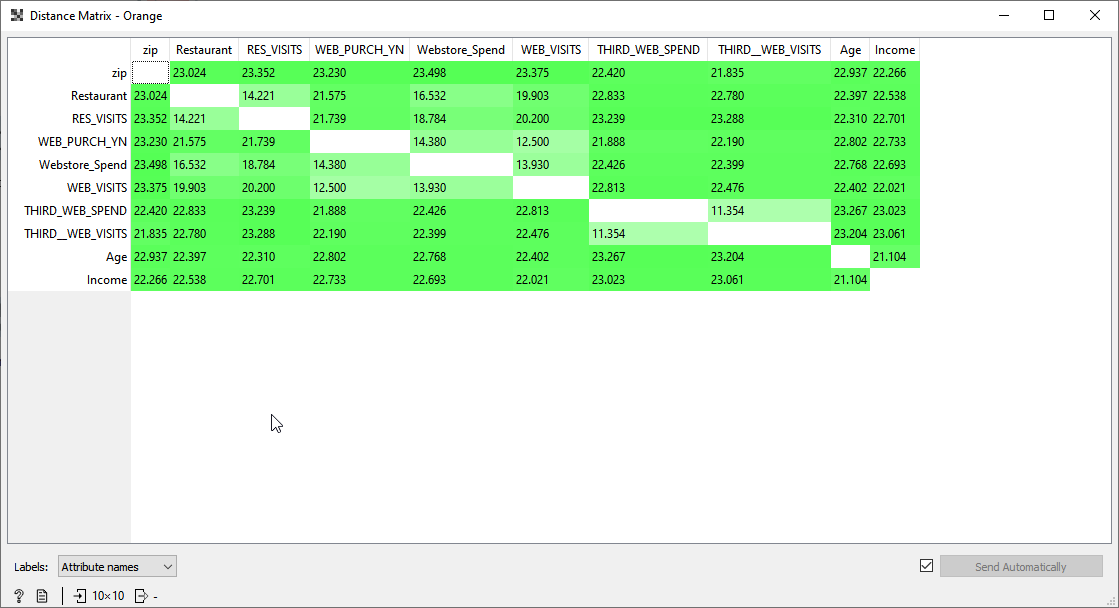
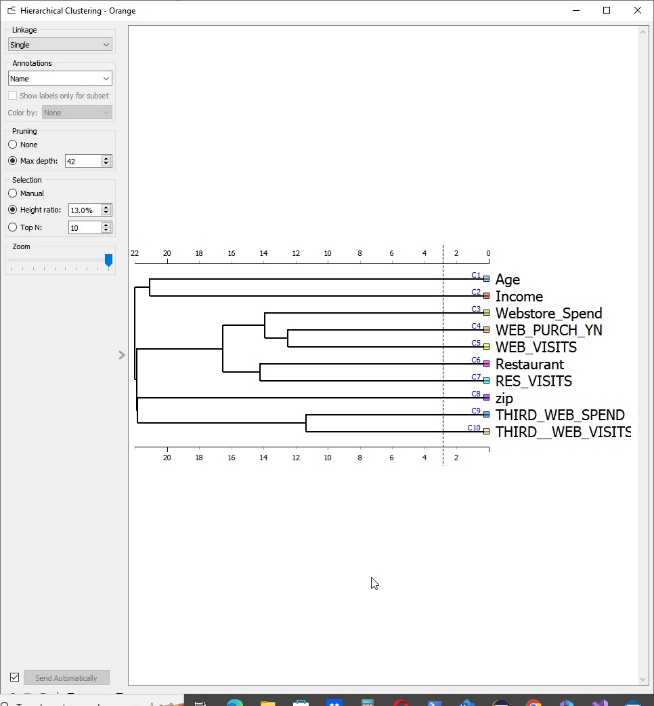
Diagram, schematic

Description automatically generatedA picture containing table

Description automatically generatedTable

Description automatically generated

Clustering from Orange provides a more concise clustering limited to 10 clusters but provides similar information as the column summary provided in JMP. Distance matrix provides insight into variables that are closest together in relation r to column summary in JMP software.



I initially created a clustering diagram using all the variables that was provided in the survey data to find clusters that would suggest what metrics best evaluated the customers and view correlations. I feel that locating these clusters would allow for further evaluation of the customer base and provide information to evaluate where the organization can look to increase their online sales. This suggested that there are 20 natural clusters in the provided data. I found this to be quite extensive and created a 1 to 1 cluster to variable ratio. This did not provide insight to the question of increasing online sales, but it did provide information of where the clusters were most significant and suggests that there is a significant reason to believe that web purchases and web spending are significant correlation variables. I then created a second cluster diagram with the variables from the column summary above that were above 50%.

Diagram

Description automatically generatedTable

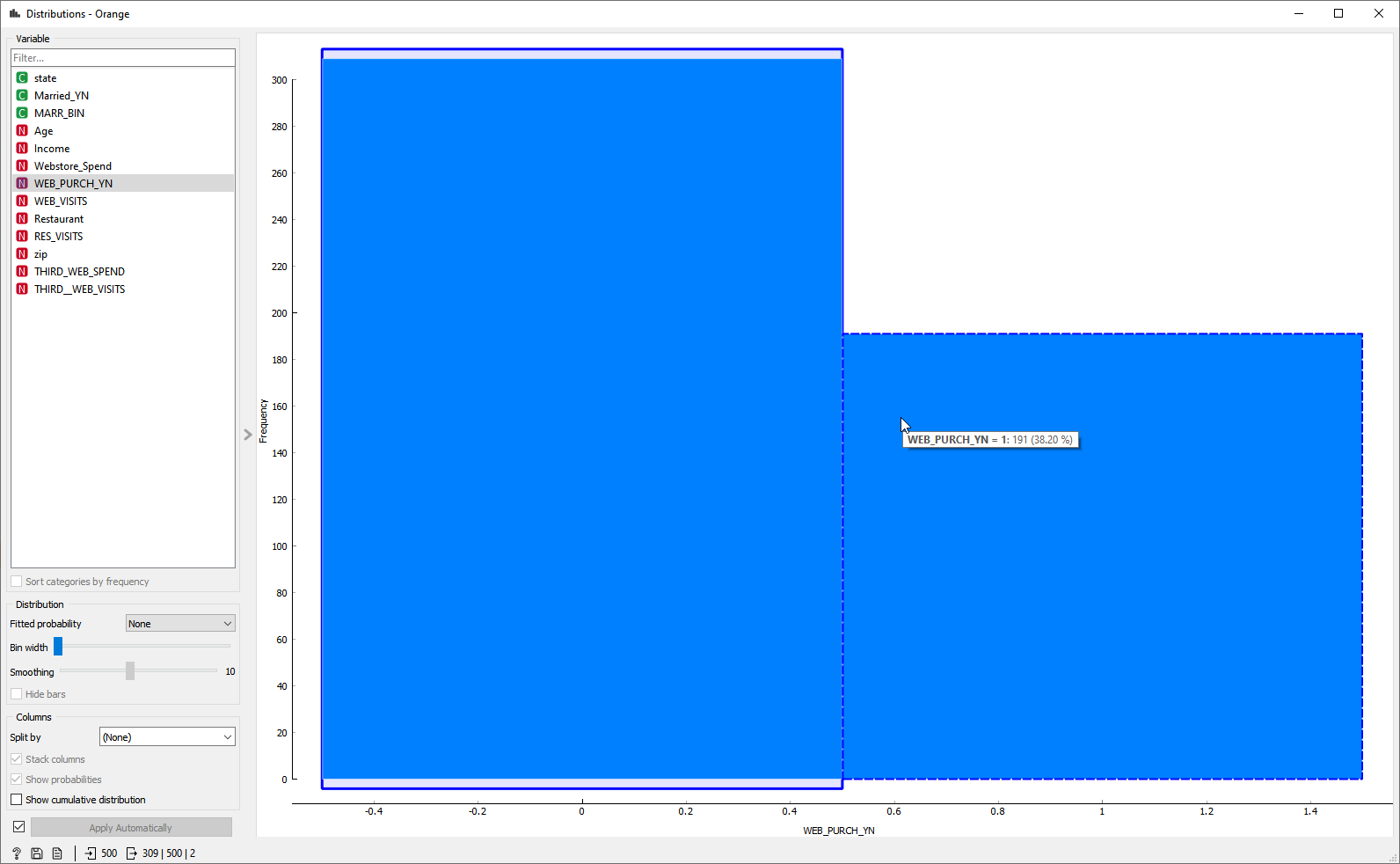
Description automatically generated

This clustering with fewer variables did not decrease the number of clusters but it did provide additional clarity to the variables that were significant and shows that all variables now accounted for more than 50% of the clustering data. I feel that this provides the clarity to be able to evaluate additional characteristics of the variables to find correlations in the data. The next step was to perform x and y comparisons as well as linear and logical regression modeling to find metrics for recommendations to increase online sales.

**WEB\_PURCH\_YN**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Freq  Share | | WEB\_PURCH\_YN | | |
| 0 | 1 | Total Responses |
| All | A | 309  61.8% | 191  38.2% | 500 |

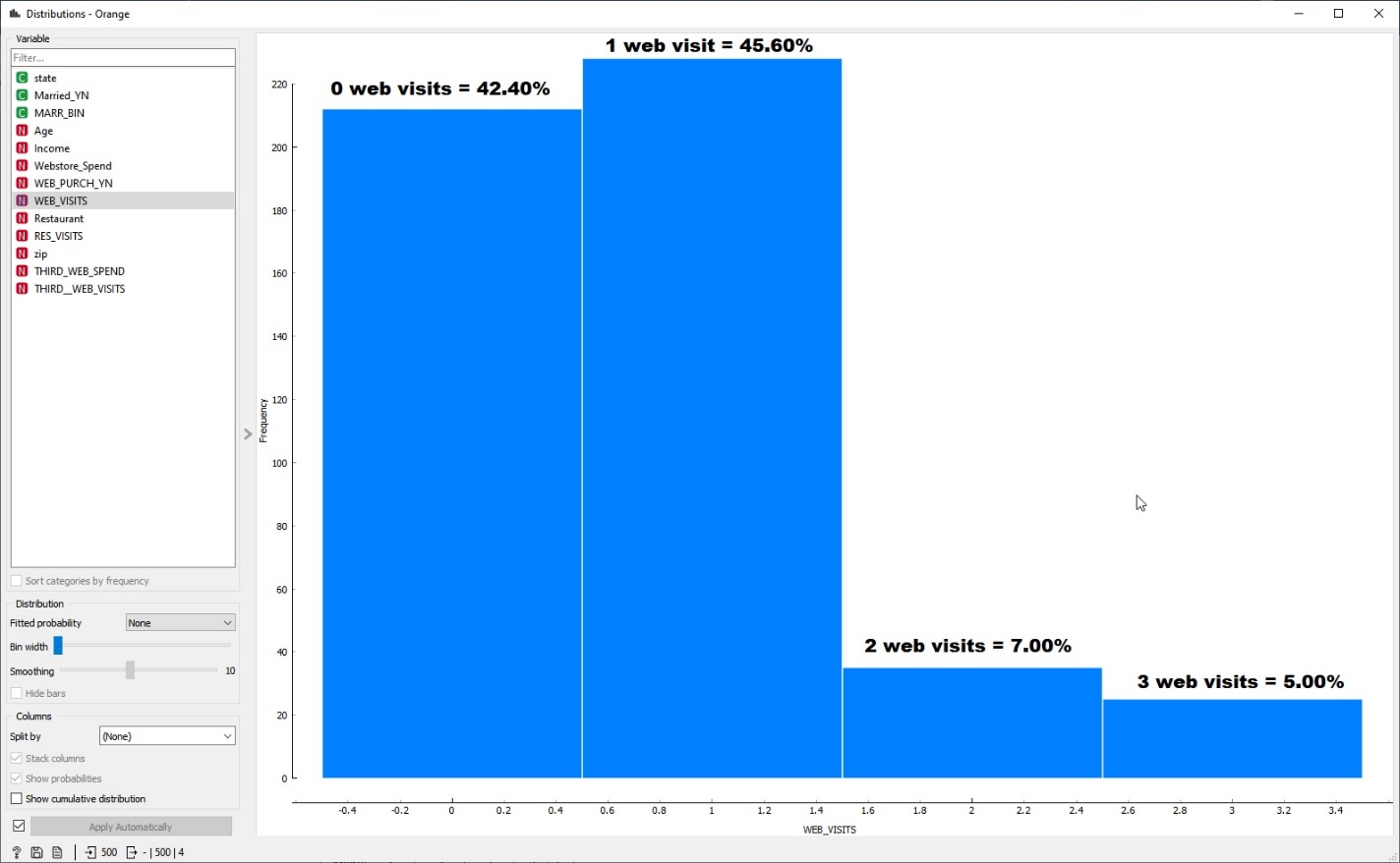
From orange



**WEB\_VISITS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Freq  Share | | WEB\_VISITS | | | | |
| 0 | 1 | 2 | 3 | Total Responses |
| All | A | 212  42.4% | 228  45.6% | 35  7.0% | 25  5.0% | 500 |

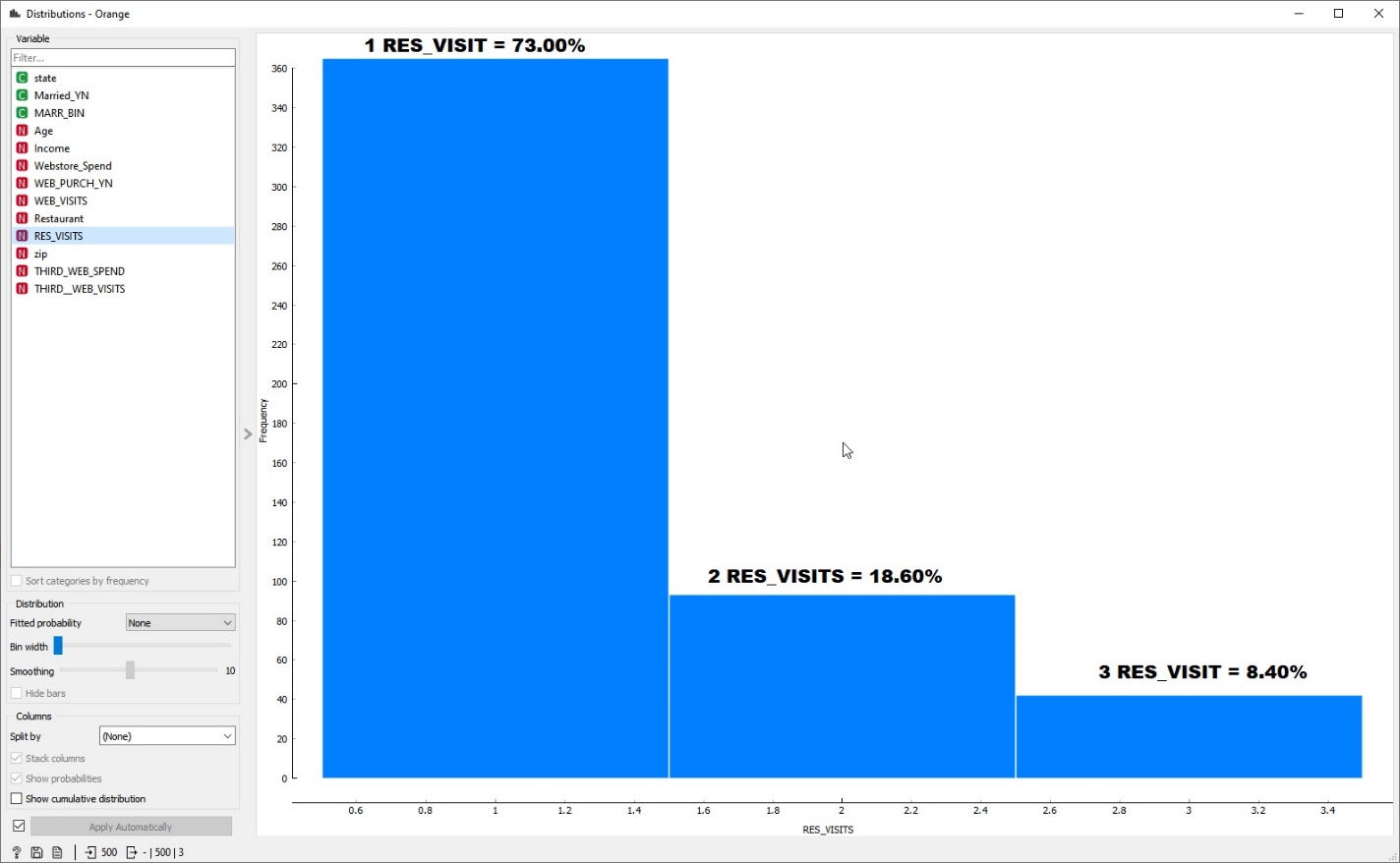
From Orange Software



**RES\_VISITS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Freq  Share | | RES\_VISITS | | | |
| 1 | 2 | 3 | Total Responses |
| All | A | 365  73.0% | 93  18.6% | 42  8.4% | 500 |

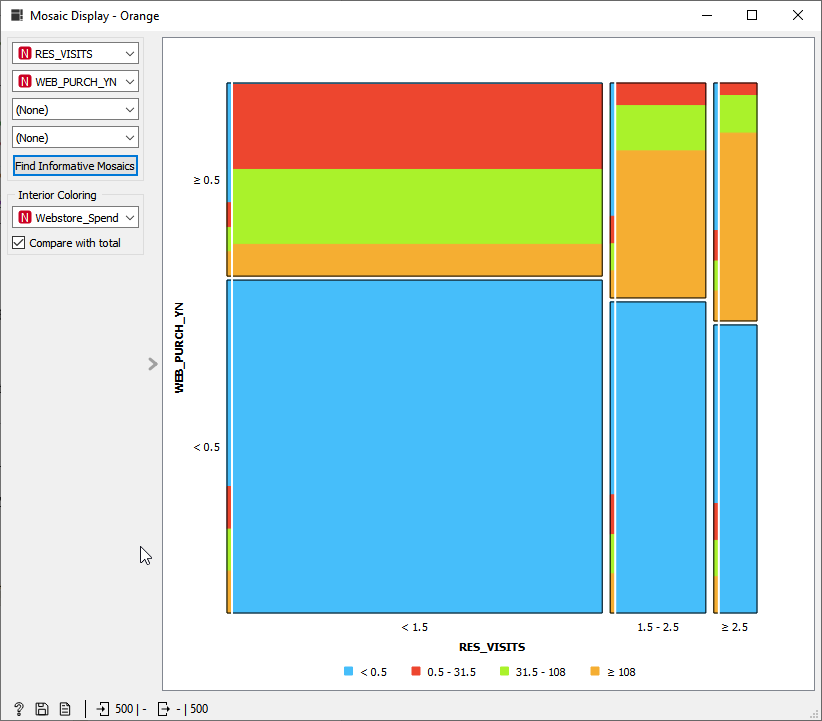
Image from Orange Software



**Partition for WEB\_PURCH\_YN**



Image from Orange software

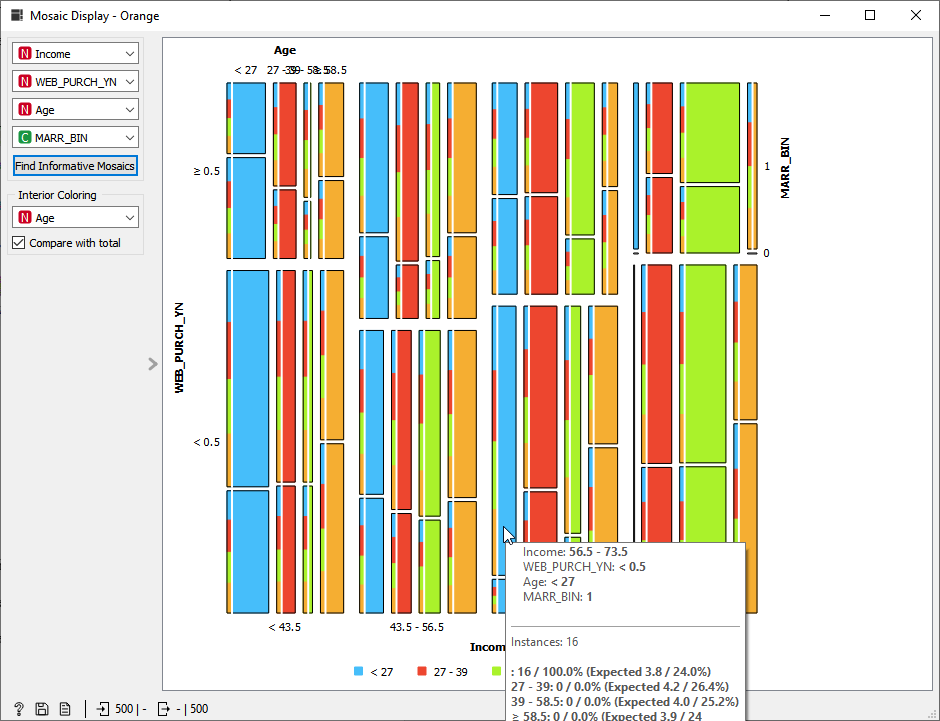
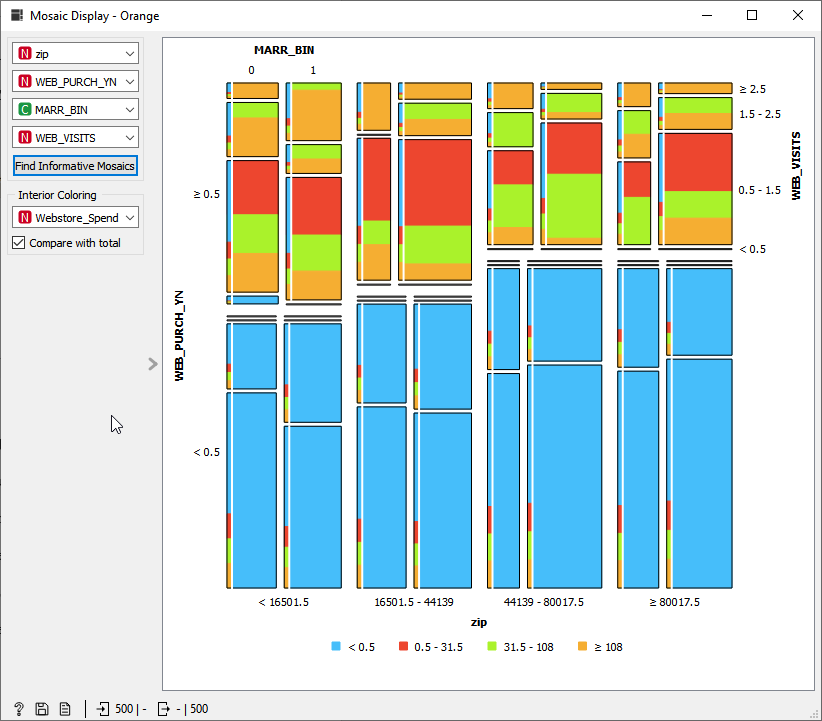


The tables here provide information for the metrics that are related to online purchasing. The first table shows that 61% of participants have not made an online purchase while 38% have. The next tables shows that 42% have not visited the online web store while 57% have visited at least once and only 13% have visited more than once. The final histogram provides an overall partition for web purchases vs restaurant visits. This graph would suggest that there is a significant number of customers that visit the restaurants but do not make web purchases. Most importantly are people who have been to the restaurant only once are the least likely to shop and spend on the web.

Graphical user interface, calendar

Description automatically generated

Images from Orange Software – I found that these visualizations were better created and annotated in the JMP software which I thought was interesting because Orange is really strong in this area on many other representations. I would have to further experiment with these to see if their appearance can be modified and improved.



The two graphs provided above provide significant insight into the groupings of the survey data. The first (above left) is web purchases yes/no by zip code and grouped by marital status and web visits. This histogram suggests that the group that is most likely to make web purchases are married individuals and multiple site visits are more likely to end in a purchase. Most notably shown in this graph is that there are no married participants that have not visited the site. The next graph (above, right) compares web purchases based on income and grouped by age and marital status. This graph further shows that married individuals over 25 years of age are the most likely to make web purchases.

## Validity, Reliability, Limitations

The data that is provided in the data sample has been shown to have some significant concerns with validity and accuracy. These concerns have been considered and will be discussed here. The data provided has two duplicates of data/inputs that were noted and removed. The duplicate entries are married\_YN and MARR\_BIN, city, county, state, and ZIP. For these responses I discarded the married\_YN using the married binary inputs instead to provide a more concise number entry in lieu of the string entry (Orange allows for more user interaction in the importing of data stage when setting up projects and I found that I could make the married YN categorical which allowed for additional manipulation). I found that the ZIP data was entered as a continuous variable which made the software treat it as a numerical value rather than a significant single number. Modified the entry type to be nominal and this seemed to correct the issue. By changing the data type for Zip I was able to eliminate the need to compare state, county, and city individually since the Zip would cover all of them. To ensure that the entries were correct I compared the city/state to the zip code and would add a script to the software to evaluate this upon data entry at the data warehouse. I have assumed here this has been completed already. Also, because the store restaurant number has little meaning outside of the organization and there is no correlation to store visited and online sales, I have left this variable out of all my evaluations. The values that could cause issues with data research and presentation are additionally found with the variables of third visits, web visits and restaurant visits. This data can be difficult to interpret because the values for third visits shows 0,1,2,3 which reads more as which visit it is or that the clients have only visited up to 3 times. The data for web visits provides the same limitation in that it only shows values as 0,1,2,3 suggesting that these are the only options for visits. Finally, the restaurant visit number is good to show that there are repeat customers, but it does not provide for which visit was the impetus for web visits and if additional visits led to additional web sales.

## Resulting Decision Influence

Given the data provided above I would recommend that the Bubba Gump co. increase their marketing to individuals that are married over the age of 25 and with an income of $40,000 per annum. This would increase sales of individual that are in the demographics of existing online customers. I feel that by targeting these areas the organization will be able to maximize sales from an existing target cluster. Additionally, I would recommend that the organization increase surveys or survey data to include website metrics and surveys for online experience. This information coupled with the customer demographic data would allow for marketing to groups outside of the areas that are currently producing online sales and allow for inspection and discovery of circumstances that are mitigating online sales outside of the target group. The data suggests that online sales are very minimal (between $10 and $450) and quite sporadic. The website should be evaluated for ease of navigation, products and services provided and time on page/site. The data also suggests that multiple visits does not equal multiple sales. This should be evaluated as well, it seems that after the initial visit, each additional visit should conclude in a sale.

## Visual Evaluation

The histograms and graphs that have been provided are a small amount of the visualizations that could be provided. I have chosen the graphs that best represent the business problem as it has been provided to increase online traffic/sales of Bubba Gump Co. website. These graphs show how the data can be reviewed and interpreted. The clustering dendrograms have also pointed out areas that the data can cause misleading results (zip and yes/no or 0/1 responses) and provide misinformation as well as inaccurate decisions as these lead to myopic interpretation of the data due to limited response possibilities.

## Next Steps

The next steps for the organization based upon this report would be to increase their marketing to the groups that are spending and are more likely to spend at the online store (married over the age of 25 and with an income of $40,000 annually). This will increase web store sales immediately. The long-term recommendation would be to create and view analytics for the online store most importantly, time on site, time per page, sales per customer and how they found the site. The organization should also increase awareness of their online store through its brick-and-mortar restaurants. The analytics suggest that there is little influence to drive customers from the restaurant to the stores. If the company is not already doing it, I would recommend providing a downloadable menu on the site that would encourage customers to visit the site to increase online presence and build online relationships with customers.

## Reference(s):

## Andrea Ahlemeyer-Stubbe, Shirley Coleman (2014). A Practical Guide to Data Mining for Business and Industry. Wiley Global Research (STMS). <https://mbsdirect.vitalsource.com/books/9781118981863>

Chen, D., Sain, S. L., & Guo, K. (2012, August 27). *Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining*. Journal of Database Marketing & Customer Strategy Management. Retrieved September 17, 2021, from <https://link.springer.com/article/10.1057/dbm.2012.17>.

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) [Orange: Data Mining Toolbox in Python](http://jmlr.org/papers/volume14/demsar13a/demsar13a.pdf), Journal of Machine Learning Research 14(Aug): 2349−2353.

Download and additional information at <https://orangedatamining.com>

Hennel, P. (n.d.). *Why Retailers Should Care About Data Mining : Intelligent Enterprise*. Https://Www.Silvon.Com/Blog/Retailers-Data-Mining/. Retrieved September 17, 2021, from https://www.silvon.com/blog/retailers-data-mining/

(n.d.), Brock, Tim. [www.displayr.com/what-is-dendrogram/](http://www.displayr.com/what-is-dendrogram/)

Images from Orange not provided in report above

General Setup – This was one of Orange Software’s most redeeming qualities with the images below and the data file that was evaluated a project can be setup by dragging Oranges widgets to the page and setting up matching the images and import settings and the project will be operating the same. This makes it modular and repeatable for many users and is great for collaboration.

