5-2 Milestone Four: Enhancement Three: Databases

Azita Dadresan

CS-499 Computer Science Capstone

Southern New Hampshire University

Professor Dr. Neil Kalinowski

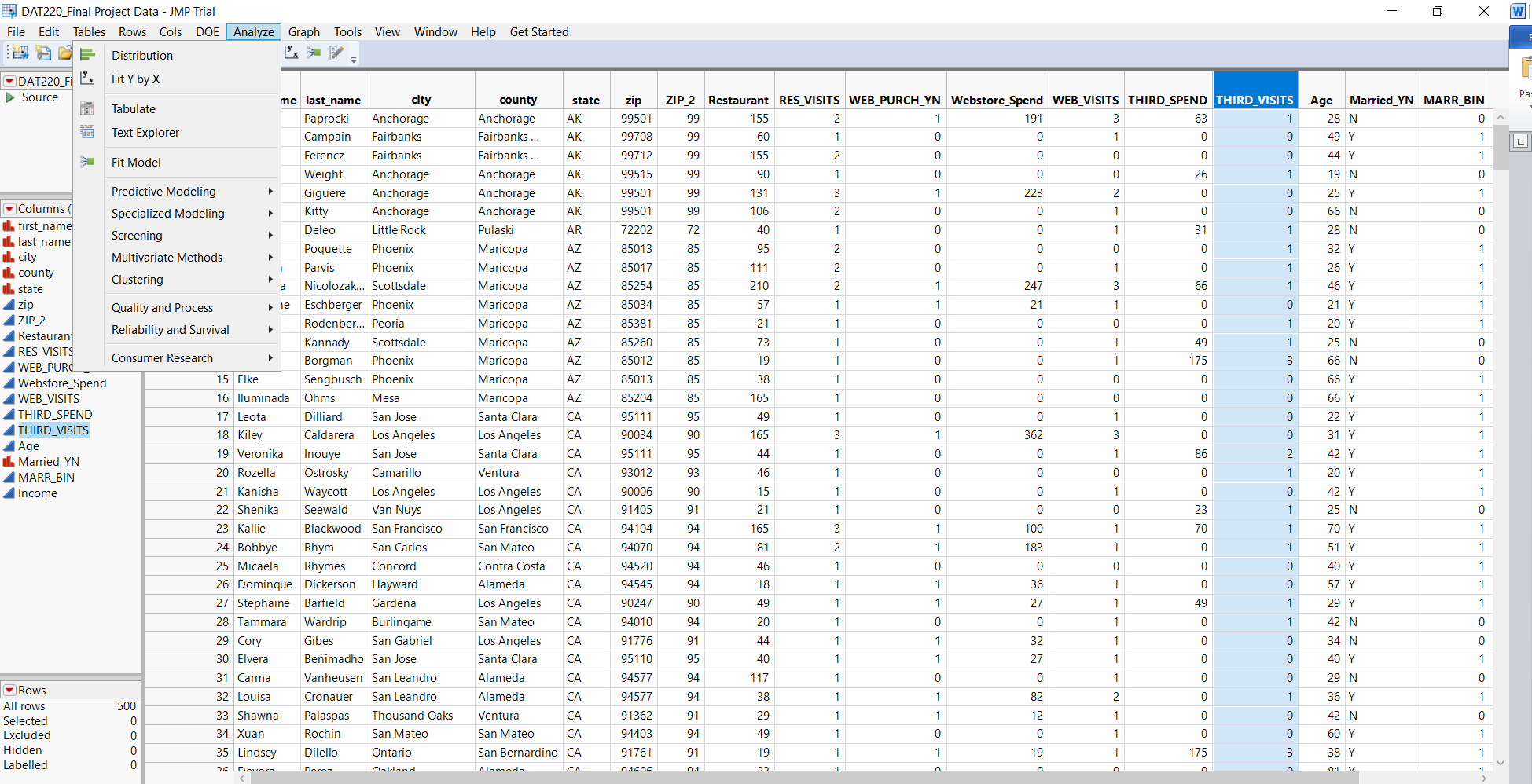
Oct. 5th , 2019

In Bubba Gump datamining DAT220, we need to find potential customers from data collected online from the bubba Gump website. We check which city they are going to buy it from, how much they spend, what ages they spend more. Our potential customer income, gender, buys from website or goes to the store. We want to know if they are first time buyer or third time buyer.

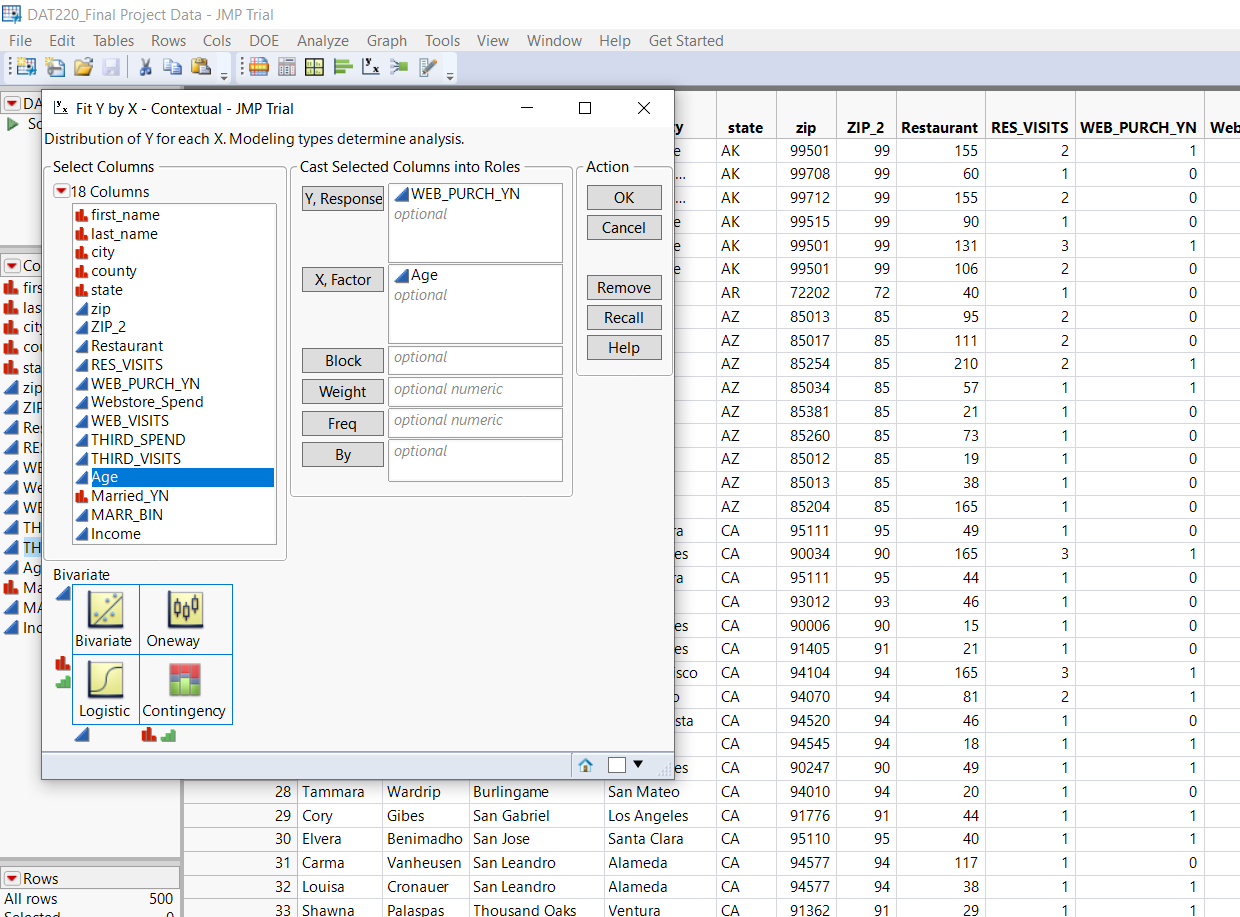
I described more information about the process of gaining the result, not just the result. Working on datamining software like JMP helps programmers to make a better decision and make scripts according to that data. It is vastly used in Artificial intelligence and Machine learning. This is the way that machine can learn how to decide to find a potential customer such as Bubba Gump analyzing data. The result is like the human brain which knows who the target customer is for advertising. It can save lots of time and budget for marketing companies to find good customers faster with lower cost.

I add more results for data analyzing for future enhancements. In Bubba Gump data I added more functions that analyze the data better. I need to get the model of data predicting what will happen based on analyzing the data.

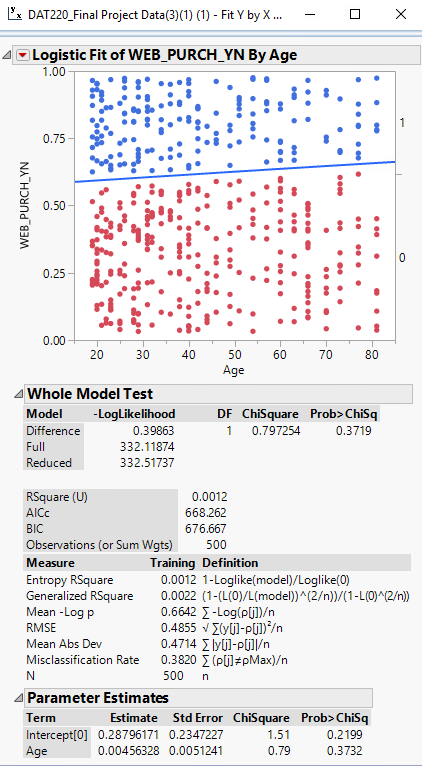
I used Regression tool, Multivariate Dependence Techniques, Categorical Data.

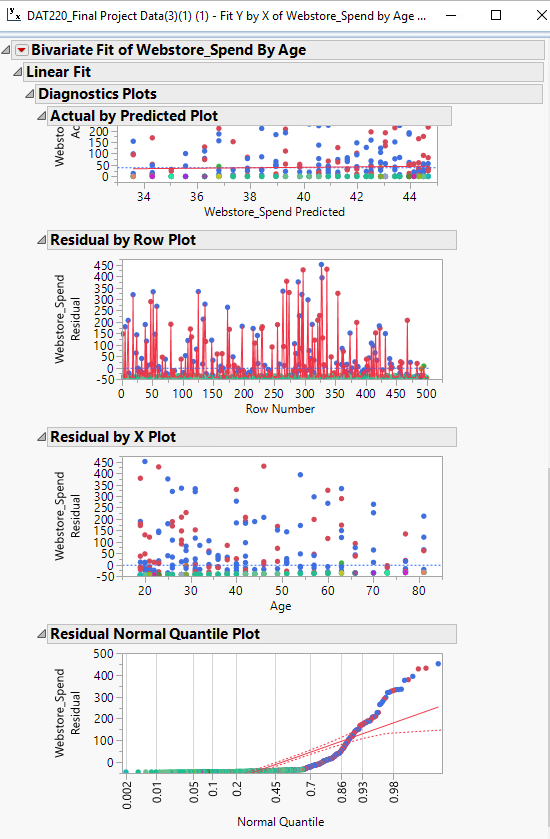


I choose web purchase yes or no as y response, and age as x factor.



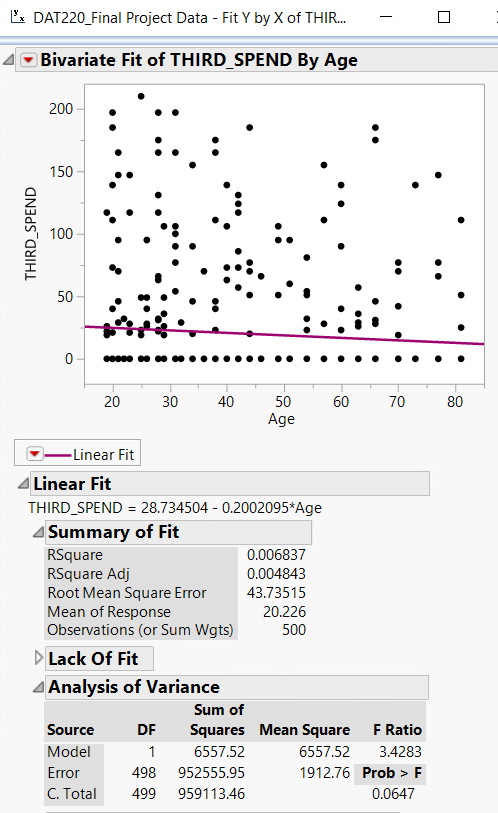
When we compare the customer’s age, they are mostly younger to purchase. Blue is purchase, Red is no purchase and X is the ages, Y is the web purchase.



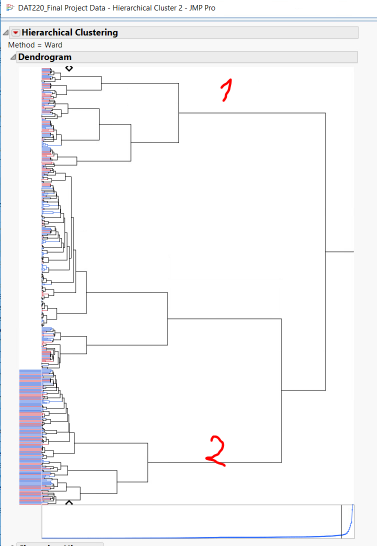


In the simple linear regression such as comparing the webstore spends on the Y and income on the X shows that web spend normally 50 for the different incomes. The incomes are between $30 to $60 who have spent on the website.

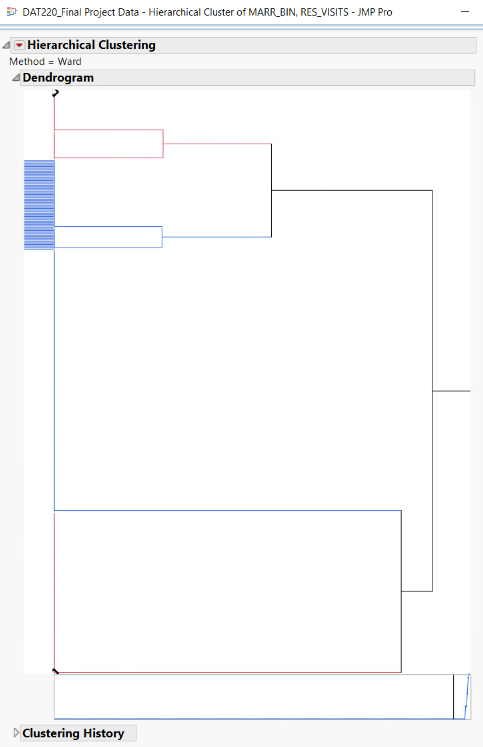
The below graph depicted the third\_spend bivariate fit by age. It shows that the most third party spends are 25 times and mostly between 20 to 30 years old.



When I compare the states, restaurant, res visits, web purchase YN. We can see that the biggest proportion had the restaurant visits only once, then no purchase at all. However, we need to focus on the parts that had restaurant visit more than one time first, then focus on the customers that had once restaurant visit, and once web purchase. I did the numbering according to the rank. First we need to focus on number one, then one number two.



Here I find association in hierarchical clustering between restaurant spend and marital status. The customers that had more than one restaurant visits was mostly married and would spend more on the third party.



Build script, share the result with interactive format.

In summary, we can see the mostly unmarried, from the states OH, NJ, MD, MA, NY, and TX, young ages, and the income between $30 to $60 per year will buy online. However, the mostly married, ages 28 to 57 years old, more than once restaurant visit, the income $44 to $74.75, had more third parties spend and from states CA, NJ, NY, FL, TX, PA, OH, MD, and cities New York, Philadelphia, Chicago, Miami, Baltimore, Gardena, Milwaukee, Orland, Phoenix, San Francisco.

We need to check the data, there were no missing data, not misspelled, no outliers, no outdated data, the information should be based on the fact, no bias and violation against the regulations, avoid incorrect values. I need to focus on the actual data not predicted data and on R-squared monitoring. In my report, each time I was getting the scattered data and not getting a good model, I was focusing on the tree clustering and association for example when the data in the web spends and third spends had lots of outliers, I used clustering to review any association and the reason that what it happened. I got a good result for them that I mentioned in the previous section.