# CS-499 Module 5 – Enhancement 3 Assessment

Database Data Mining

Erik Wilhelm

erik.wilhelm@snhu.edu

Southern New Hampshire University

Briefly describe the artifact. What is it? When was it created?

The artifact that I selected for this enhancement is from DAT-220 Fundamentals of Data Mining that I took in September of 2021. The artifact is from my final project in the course. The project was to use a data file (DAT220 Final Project Data.csv) to perform Data Mining evaluations and provide a fictious Restaurant company with results that could be leveraged to increase their online sales (also referred to as web sales, WEB\_SPEND and WEBSTORE\_SPEND). The data file was created based on customer surveys and feedback from a national franchise restaurant. The data provide various categorical and numerical data points for evaluation. The data was evaluated using JMP statistical software, this software was provided by SNHU technology. I no longer have access to the software so I can only refer to my records from the class and reports submitted therin.

Justify the inclusion of the artifact in your ePortfolio. Why did you select this item? What specific components of the artifact showcase your skills and abilities in software development? How was the artifact improved??

Data Mining is a valuable tool that can be employed to provide predictions, enact change, and to clarify unclear information. I took data mining as a requirement of my Computer Science program, but I quickly found that the information was not only very applicable in my work as a project manager but extremely interesting and enjoyable. Data mining relies on the creation and storage of databases so it can relate to virtually anything worthy of creating a list. Having an understanding and the ability to develop a database is important for creating databases capable of filling a specific need but using them for information and to learn there are extensive online resources for sample and test databases.

I chose to explore data mining for this artifact because the experience and understanding that I will develop from this enhancement will enhance my productivity and profitability in my career. For this enhancement I have used the (DAT220 Final Project Data.csv) dataset that was provided in DAT-220 and evaluated it using Orange Data Mining Software. Orange is developed by [Bioinformatics Lab](http://www.biolab.si/) at University of Ljubljana, Slovenia, in collaboration with the [open source community](https://github.com/biolab/orange3) (<https://orangedatamining.com/citation/>). Once I found that I had an interest in continuing my learning and experience with data mining I spoke with my instructor and he told me about Orange being a very powerful opensource datamining software project that had a lot of support in both enterprise and educational uses. I finished the class and did not revisit Orange or the practice of data mining until this project.

The plan that for this enhancement was to setup Orange Data mining using the most recent release to ensure that I have the most up to date features secure release. I reviewed the data set that I was provided in DAT-220 using Orange and compared my findings to those that were provided in my original report. For this artifact I will submit a working data mining file that is native to Orange Data Mining, images from my evaluations, and assessment of findings. The skills that this project enhancement will showcase are the ability to employ past and existing experience with Databases and Data Mining in order to learn new software and perform tasks efficiently. I will showcase my ability to import, work with, and evaluate databases using Data Mining theory and open-source data mining software. This enhancement improves the original project by importing the database information into an open-source software with the functionality of expensive enterprise software. This enhancement will make the data and visualizations accessible to anyone willing to take the time to download and experiment with new software. This enhance provides a new way to visualize database information that makes it easier to find trends, commonalities, and falsehoods in database datasets.

Did you meet the course objectives you planned to meet with this enhancement in Module Five? Do you have any updates to your outcome-coverage plans?

The course objectives that I met with this enhancement are:

Employ strategies for building collaborative environments that enable diverse audiences to support organizational decision making in the field of computer science. This has been made evident by employing the use of creating a project enhancement plan, video code review, and artifact assessment. I have provided links to information regarding Orange and its documentation.

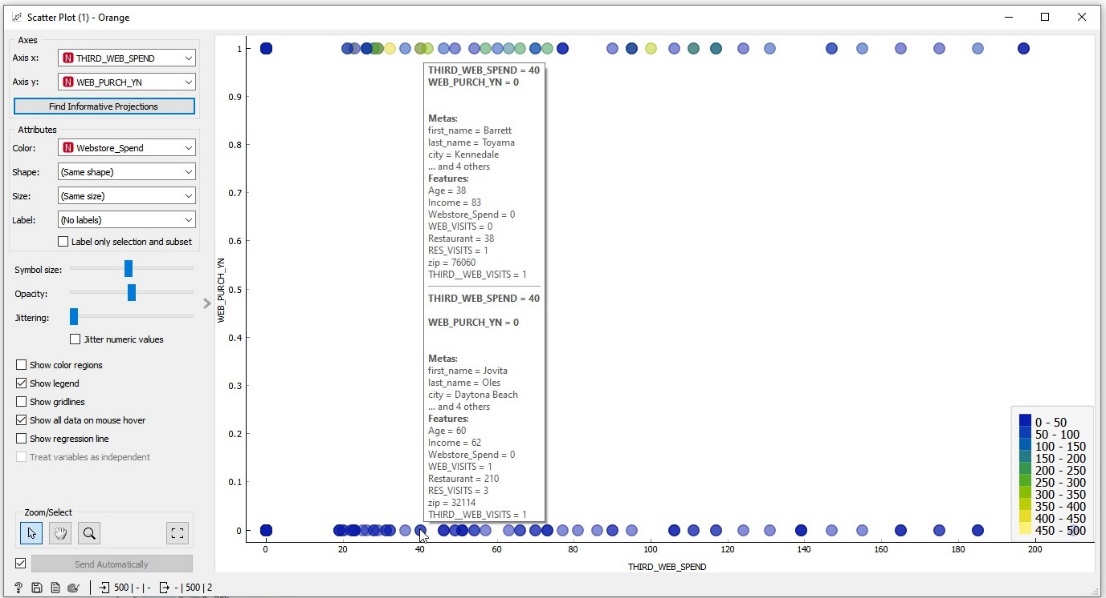
Demonstrate an ability to use well-founded and innovative techniques, skills, and tools in computing practices for the purpose of implementing computer solutions that deliver value and accomplish industry-specific goals (software engineering/design/database). This has been demonstrated by employing current open-source data mining software to evaluate a known dataset with intentions to evaluate it as a work applicable software that can provide insight into project outcomes and efficiency. I wanted to use previously unfamiliar software that is free and opensource to evaluate a dataset that I previously worked with using enterprise paid license software. This is important because I work in a small family business with a limited software budget that is allocated for day-to-day required software. The ability to use opensource software to evaluate our bidding reports and project financials to increase efficiency in bidding, project management, and lead to overall company profit.

Develop a security mindset that anticipates adversarial exploits in software architecture and designs to expose potential vulnerabilities, mitigate design flaws, and ensure privacy and enhanced security of data and resources. This enhancement supports a security mindset by ensuring the software is up to date and works with computer operating system. Orange Data Mining has increased the security of their software be employing Virtual Environment for each project created using its python coding base. Orange Data Mining is open source and community supported this can lead to significant building of software support but can potentially be open for bugs. Updates are not automatic and require permission prior to installing so this can mitigate corrupted or untested software. I found several inaccuracies in the data both in the initial report and in the evaluation using Orange. While learning data mining principles in DAT-220, I learned for the security of the information and one’s own credibility and a data scientist it is imperative to point out areas of possible inaccuracies in their data and interpretation of the data. In an effort to keep this assessment brief I have linked my original report which contains images and updated interpretations from my Orange date review.

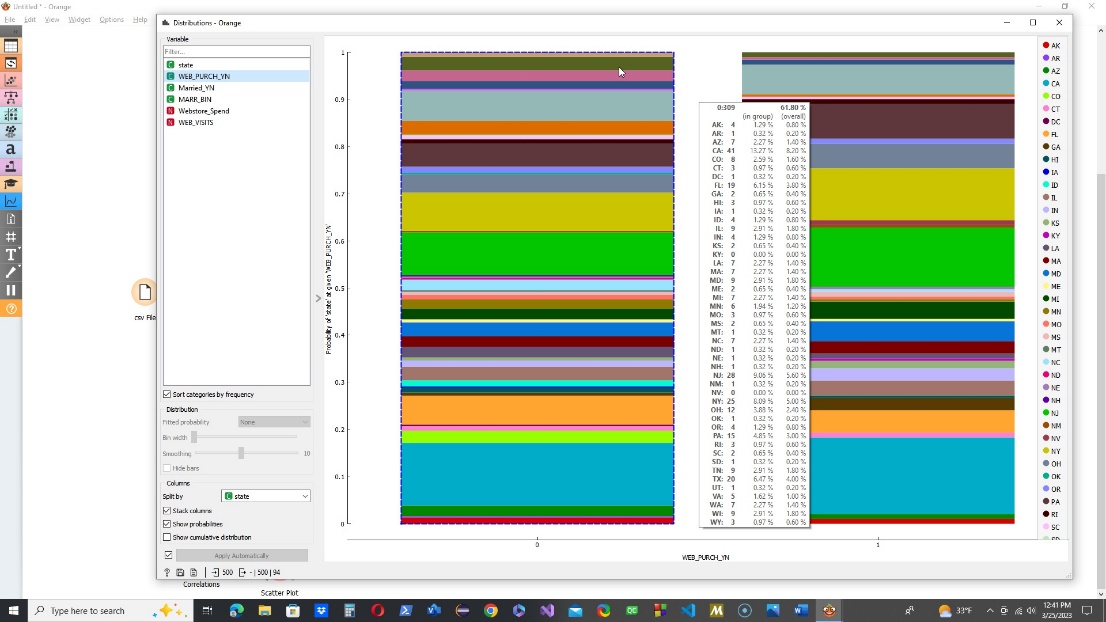
Reflect on the process of enhancing and/or modifying the artifact. What did you learn as you were creating it and improving it? What challenges did you face?

I started the project with a clear direction as to the path I thought it would take. I found that as I used the Orange software, I tried to mimic the JMP tests and outcomes. I started my evaluation of the dataset over multiple times while working on this artifact. Each time I tried to mimic my work from my initial project, the Orange software lead me down multiple different paths to find similar outcomes that I had previously. Orange helped me to visualize findings that I misinterpreted previously as well as find information that did not answer the initial business question of finding ways to increase web traffic and sales but were interesting. When performing my initial evaluation there was a variable THIRD\_SPEND and at the time I thought that was for third time visiting the webstore spent money. While working in Orange, which is a very visually positive software I found that there were entries in the database that had zero web visits but values in THIRD\_SPEND, this led me to believe now that it was third time at a restaurant. This was an important finding because in the first report I took that as a strong metric to indicate web store traffic and spending and this greatly affected my interpretation of the data in the less visual JMP software.

Seen in image below from my Orange file

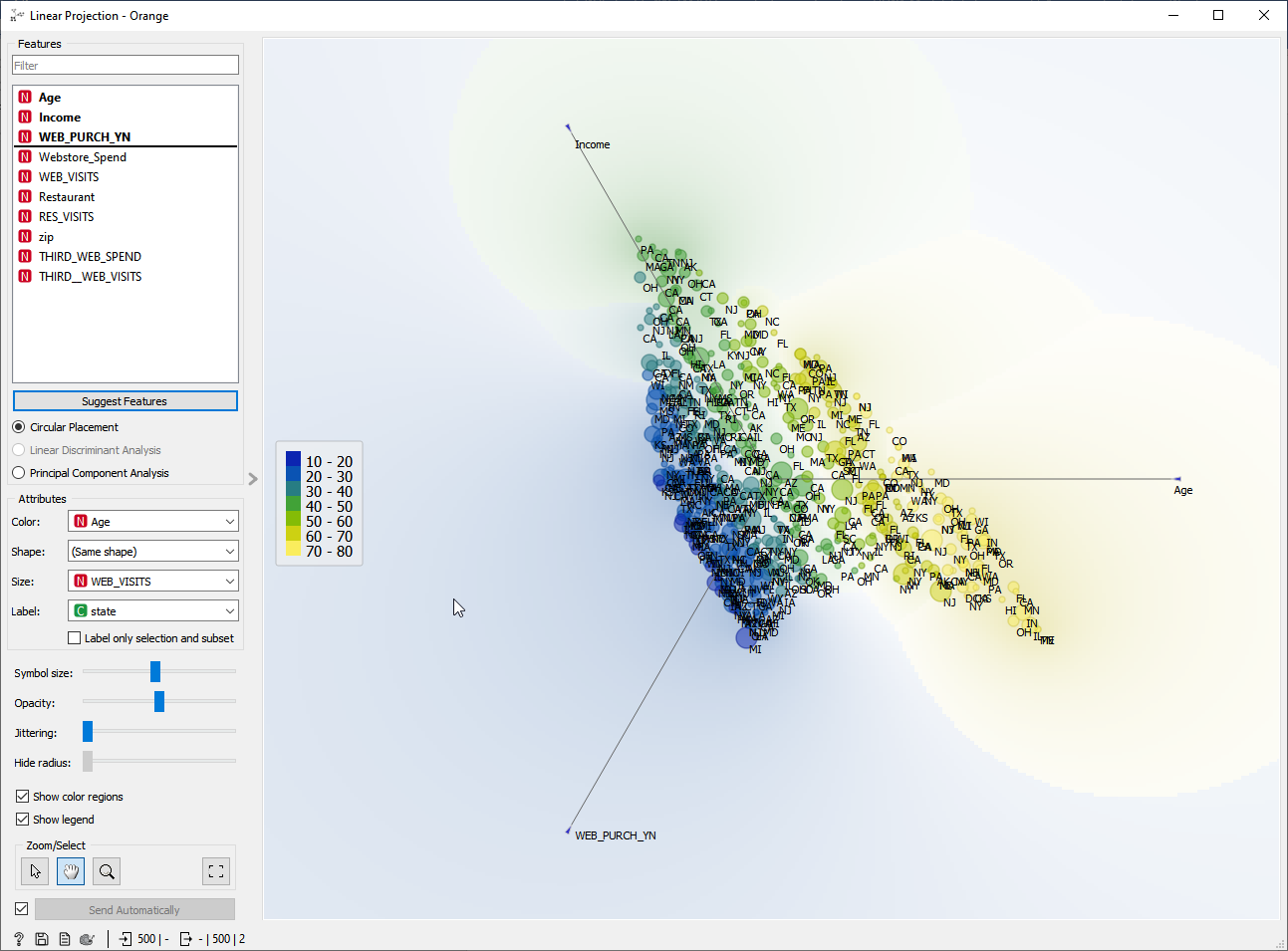


I did find that my initial evaluation to be common in both programs below is a visualization from orange showing most common states for web purchase was New York, New Jersey, and California as found in my report using JMP.

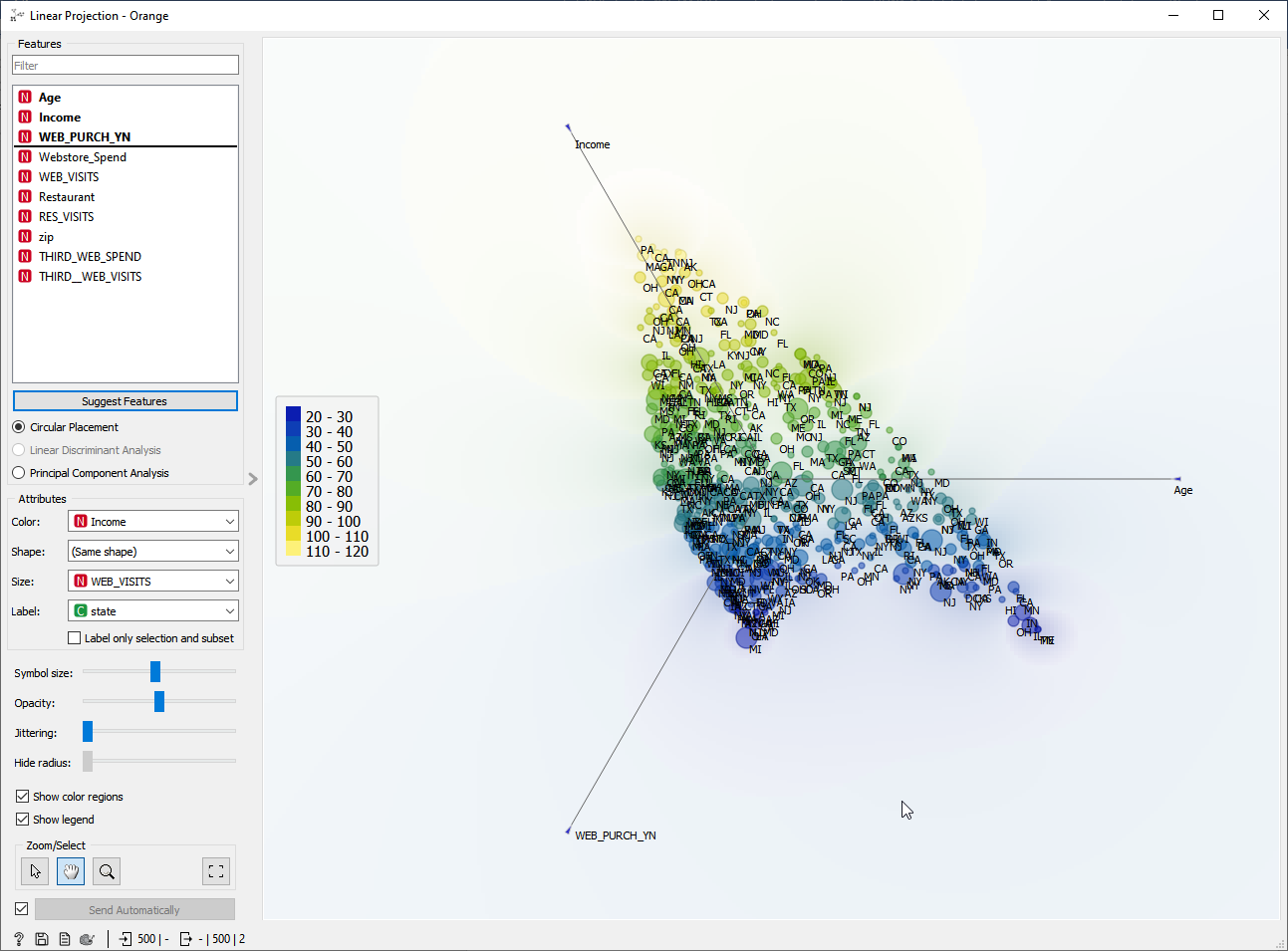


Using linear projection to evaluate multiple variables and how they might relate to each other confirms my initial evaluation that the predominant web sales are to patrons in the ages of 20-40 with an annual income of 40-60k per year.

Colored to show age



Colored to show income



References:

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. Link to Orange Software Website <https://orangedatamining.com/>

Erik Wilhelm - DAT-220-T1137-21EW1-Module 8 - Final Project\_10\_24\_2021-Revised\_03\_26\_2023