Visualizing Customer Segmentation in 3D using Paraview

Jongwon Lee
Indiana University
leejojo@iu.edu

Abstract

We describe a series of trials on visualizing information data on Paraview which is mainly used for scientific visualization of 3D objects such as geographic, microscopic, architectural, etc. With Information Visualizations done in 3D space, we aim to derive insights focusing on determining which factors are significant on the topic of customer segmentation. According to the analysis of visualization, we found Ever_Married, Age, Graduated, Profession, Spending_Score were significant for customer segmentation and Gender, Work_Experience, and Family_Size were less significant. In addition, we identify challenges related to limited data types, exporting axis and legend and displaying the visualization with interactivity.

Keywords— Scientific Visualization, Information Visualization, Customer Segmentation, Classification

1 Overview

While Scientific Visualization is used for displaying relatively known problems, Information Visualization is used for deriving hidden insights that are difficult to be discovered without visualization.

Customer Segmentation is the technique of dividing a customer population into groups that share certain features to optimize marketing strategy. Features commonly considered for customer segmentation are age, gender, interests, spending scores, etc.

Visualizing data itself can derive insights that is difficult to be discovered from numerical analysis. The main goal of Customer Segmentation is to segment customers to target the right customers and come up with an optimal marketing plan.

However, machine learning results occasionally lack context and insight to the majority of the non-machine-learning-specialist audiences.

The Customer research area is the fusion of experts in various areas such as economists, marketers and data scientists. Economists and marketers are usually not familiar with understanding complex data or the analysis of the data. Data Scientist providing effective visualization along with the numerical analysis will help collaborate with economists and marketers.

The machine learning analyses and results will be the basis and guideline for my visualization. In other words, my visualization's goal is to help the machine learning results to stand out and make them easy to be comprehended by general audiences.

2 Intended Audiences

This project is for the audiences that have background knowledge in economics or that are interested in investigating customer behavior and over 18 years old. Also, it will help machine learning experts understand their data more deeply. Lastly, general Public with basic economic concepts will be able to understand economical insights derived from visualizations.

3 Data

Our dataset comes from the 2020 AV - Janatahack Customer Segmentation hackathon (Janatahack, 2020). This visualization project will be using the train dataset that consists of 8068 customers.

This dataset assumes the following scenario: An automobile company researched the behavior of 8068 customers and found that the future customers will share the traits of the previous ones. In the previous market, the researchers classified customers into 4 segments (A, B, C, D). Then, they applied different marketing strategies on 4 segments. Since this strategy was successful for

them, they plan to use the same strategy on the new market that consists of 2627 new potential customers. The hackathon thus provides information of 8068 customers (train) labeled with either (A, B, C, D) segments and information of 2627 (test) new potential customers. The hackathon expected programmers to come up with supervised machine learning methods by training a model using the train data and segment the test data.

The initial train dataset provided in the hackathon is provided in Figure 1.

	ID	Age	Work_Experience	Family_Size
count	8068.000000	8068.000000	7239.000000	7733.000000
mean	463479.214551	43.466906	2.641663	2.850123
std	2595.381232	16.711696	3.406763	1.531413
min	458982.000000	18.000000	0.000000	1.000000
25%	461240.750000	30.000000	0.000000	2.000000
50%	463472.500000	40.000000	1.000000	3.000000
75%	465744.250000	53.000000	4.000000	4.000000
max	467974.000000	89.000000	14.000000	9.000000

	Gender	Ever_Married	Graduated	Profession	Spending_Score	Var_1	Segmentation
count	8068	7928	7990	7944	8068	7992	8068
unique	2	2	2	9	3	7	4
top	Male	Yes	Yes	Artist	Low	Cat_6	D
freq	4417	4643	4968	2516	4878	5238	2268

Figure 1: Initial Dataset

Names of variables are self-explanatory except for "Var_1". "Var_1" is just an Anonymised Category for the customer and irrelevant to the data analysis, thus removed. Rows with columns that were missing data were removed. Categorical Data such as Gender, Ever_Married, Graduated, Spending_Score were converted to a numerical data for better analysis and visualization. Numerical Data were Jittered to accomplish a visualization result that data points are not overlapped.

Dataset after preprocessing had 6718 rows. (Figure 2)

For visualization, we initially decided Age(X), Spending_Score(Y), Work_Experience(Z) as variables for three dimensional axis because other variables such as Gender, Ever_Married, Graduated, Profession were categorical Data. Family_Size was visualized with the size of data point.

4 Workflow and Tools

Python and Jupyter Notebook was used to preprocess data and explore winning machine learning results of (Janatahack, 2020).

Paraview was used to create 3D visualization. Most of the visualization analysis were done in Paraview.

	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Segmentation
0	462809	0	0	22.005218	0	Healthcare	0.824661	1.081320	3.910278	D
2	466315	1	1	67.096430	1	Engineer	0.940159	0.957967	0.940883	B
3	461735	0	1	66.858022	1	Lawyer	0.013105	3.061772	1.869699	B
5	461319	0	1	55.976516	0	Artist	0.148689	1.831747	2.261938	C
6	460156	0	0	31.931783	1	Healthcare	1.038835	0.911262	2.967078	C
...	...	...	...	...	...	...	...	...	...	...
8062	463002	0	1	40.941618	1	Artist	-0.141647	2.935496	4.848366	B
8064	464685	0	0	35.052352	0	Executive	2.967443	0.999410	3.958836	D
8065	465406	1	0	33.068748	1	Healthcare	0.993703	1.138781	1.113315	D
8066	467299	1	0	26.974122	1	Healthcare	0.856663	0.957823	4.147065	B
8067	461879	0	1	36.892598	1	Executive	-0.066992	1.858047	2.921896	B

6718 rows × 10 columns

	ID	Gender	Ever_Married	Age	Graduated	Work_Experience	Spending_Score	Family_Size
count	6718.000000	6718.000000	6718.000000	6718.000000	6718.000000	6718.000000	6718.000000	6718.000000
mean	463516.571152	0.448943	0.591694	43.523023	0.636797	2.630395	1.553099	2.839689
std	2566.017254	0.497423	0.491557	16.514529	0.480959	3.405650	0.747070	1.524866
min	458982.000000	0.000000	0.000000	17.705852	0.000000	-0.318088	0.690282	0.610488
25%	461347.250000	0.000000	0.000000	30.909201	0.000000	0.077130	0.981648	1.915825
50%	463566.000000	0.000000	1.000000	40.903597	1.000000	1.013193	1.098551	2.307635
75%	465739.750000	1.000000	1.000000	52.923907	1.000000	4.086435	2.032148	3.951861
max	467974.000000	1.000000	1.000000	89.285815	1.000000	14.224213	3.316878	9.220449

	Profession	Segmentation
count	6718	6718
unique	9	4
top	Artist	D
freq	2211	1772

Figure 2: Dataset after Preprocessing

Sketchfab was used to export 3D model generated in Paraview. However, Sketchfab would not properly export the axis and legend from Paraview. Information Visualization is generally more difficult to understand without axis names, values and legends because it is then just group of data points. One way to cover this issue was to add annotation on Sketchfab. Ideally, if Paraview provides own way to export the scene to an interactive format that can be viewed on webs, it will perfectly solve this issue. Sketchfab Model links are provided for important figures in the Results section.

Collectome was used to embed Sketchfab models and presentation on a unified view.

5 Prior Visualization Work

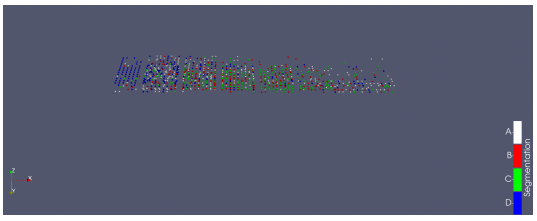

Figure 3: Prior Visualization
X: Age, Y: Spending Score, Z: Work Experience, Color: Segmentation, Size: Family Size

Figure 3 was a visualization achieved at an initial attempt. Most importantly, Jittering (spreading data points by adding noise) needed to be applied to avoid overlap of data points. Y and Z Axis needed to be scaled since it was overly crowded. Spending_Score was scaled from (1,2,3)

to (5,10,15). Work_Experience was scaled from (0,1,2,...,14) to (0,2,4,...,28). Family_Size was added as another variable by the size of data points. Better color choices for target, in this case, Segmentation, and background were needed to improve overall readability of visualization.

6 Results

The improved visualization revealed many relationships between variables and insights of customer segmentation.

Age and Spending Score were significant factor for customer segmentation since the data points cluster accordingly.

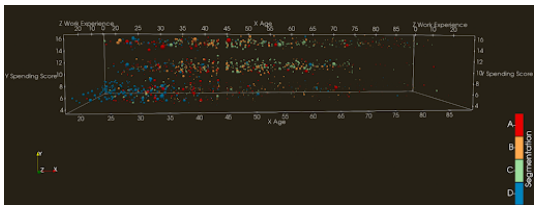

Figure 4: Segment Visualization 1
X: Age, Y: Spending Score, Z: Work Experience,
Color: Segmentation, Size: Family Size
Sketchfab Model: https://skfb.ly/orHzI

Figure 4 reveals that four segments had traits:
A: Mid Age, Low Spending Score
B: Mid Age, High Spending Score
C: High Age, High Spending Score
D: Low Age, Low Spending Score

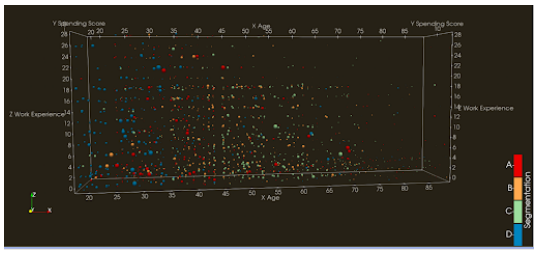

Figure 5: Segment Visualization 2
X: Age, Y: Spending Score, Z: Work Experience,
Color: Segmentation, Size: Family Size

Figure 5 addresses that Work_Experience has a mediocre impact on segmenting customers. Thus, Z axis will be replaced to another variable in other visualizations that will be introduced later. However, Figure 4 nor Figure 5 cannot justify Family_Size plays significant role on customer Segmentation.

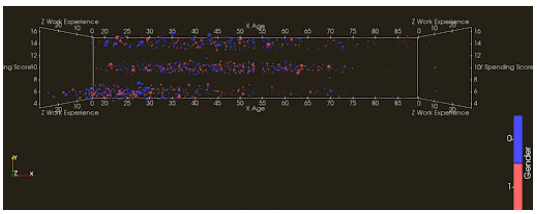

Figure 6: Gender Visualization
X: Age, Y: Spending Score, Z: Work Experience,
Color: Gender, Size: Family Size

Figure 6 shows Gender does not play an important role on customer segmentation.

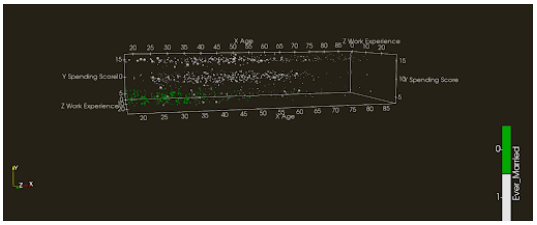

Figure 7: Ever Married Visualization
X: Age, Y: Spending Score, Z: Work Experience,
Color: Ever Married, Size: Family Size

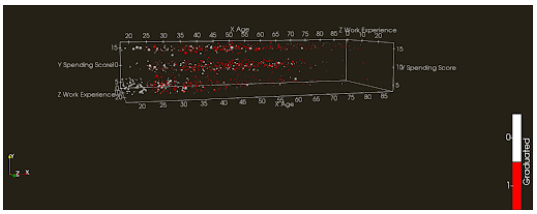

Figure 8: Graduated Visualization
X: Age, Y: Spending Score, Z: Work Experience,
Color: Graduated, Size: Family Size
Sketchfab Model: https://skfb.ly/orIs9

Figure 7 and 8 show Ever_Married and Graduated are crucial for customer segmentation.

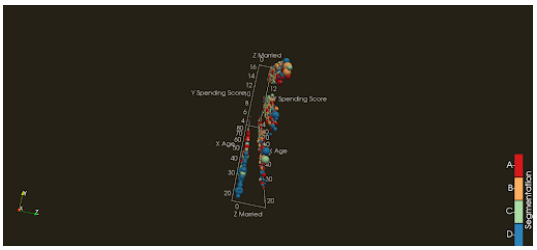

Figure 9: Ever Married on Z Visualization
X: Age, Y: Spending Score, Z: Ever Married,
Color: Segmentation, Size: Family Size
Sketchfab Model: https://skfb.ly/orI9E

Since Figure 7 revealed that Ever_Married effectively segments customers, we decided to substitute Z axis with Ever_Married which was previously Work_Experience. Figure 9 justified that the variable Ever_Married segments customers well. For instance, Segment A and D were mostly not Married and B and C were mostly Married.

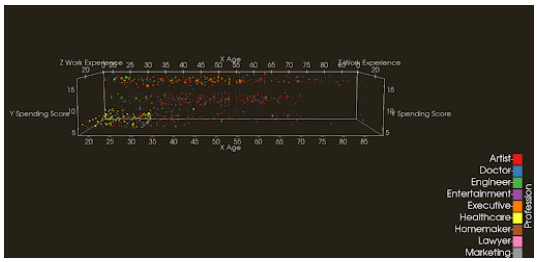

Figure 10: Profession Visualization
X: Age, Y: Spending Score, Z: Ever Married,
Color: Profession, Size: Family Size
Sketchfab Model: https://skfb.ly/orIrY

Figure 10 shows how professions are also segmented mainly by Age and Spending Score.

Professions are segmented into 3 groups by:
Spending Score (Low to High):
(Healthcare, Entertainment) - (Engineer, Artist) - (Executive, Lawyer)
Age (Low to High):
(Healthcare, Engineer) - (Executive) - (Lawyer).

Finally, Profession and Segmentation was not visualized together because those two were both categorical variables. This is because the order of the data is not meaningful and thus meaningless to put one of them on an axis. Still, some insights can be derived by comparing Figure 4 and Figure 10. For example, Figure 4's Segment D which had traits of low Age and high Spending_Score overlaps with the Yellow points (Healthcare Professions) of Figure 10.

Machine learning approaches such as one-hot-encoding is a standard solution for this kind of analysis but it is tricky to accurately analyze this with just visualizations.

7 Conclusion

Customer Segmentation can be significantly solved with visualization in 3D. The strength of visualization data analysis compared to standard data analysis methods such as machine learning is that a broader audience can understand the analysis. In addition, visualization analysis can derive insights that may be hidden or hard to discover in a standard data analysis. With the visualization on the dataset (Janatahack, 2020), we could identify Ever_Married, Age, Graduated, Profession, Spending_Score were significant for customer segmentation and Gender, Work_Experience, and Family_Size were less significant. Data Scientists should be able to convince marketers and decision makers better by providing 3D visualizations like the ones that are introduced in this project.

References

Janatahack. 2020. Janatahack: Customer segmentation.