Data Coverage Toolkit for Multi-dimensional Data

Akshay Parashar, Mohit Loganathan, Sujitha Perumal
University of Wisconsin-Madison
{aparashar, loganathan3, sperumal}@wisc.edu

Abstract

Understanding how the multi-dimensional data is distributed, displayed via visualizations and visually perceived is a challenging task. As computers become more powerful, there is a growing demand for tools that can help us understand and analyze multi-dimensional data distributions and how it can be effectively organized and grouped into different categories. This report presents an overview of the tool developed to create interactive visualizations to present the distribution of data using a mix of user-defined and pre-defined binning techniques. The tool involves generating mult-perspectives of the data, each of which highlights different aspects or grouping distributions of the data. The views are interactively linked and combined to provide more powerful and comprehensive data visualizations. We examine some promising design principles and techniques to create visualizations that aims at exploring data distribution across groupings by different variables. The report also evaluates/critiques upon the different viz techniques developed to build upon better designs and potential areas for future development.

Keywords: Data visualization, interaction, exploratory analysis, data binning, equal-width binning, filtering,

1 Motivation

"The greatest value of a picture is when it forces us to notice what we never expected to see."

As data becomes increasingly complex and multi-dimensional, there is a growing need for effective tools and techniques for visualizing and understanding this data. By visualizing multi-dimensional data in a clear and intuitive way, researchers, analysts, and decision-makers can more easily extract insights and make informed decisions based on the data. Thus, the increasing use and importance of multi-dimensional data in fields such as business, science, and technology has driven the need for better tools and approaches for visualizing and understanding of this data. Moreover, the users of such datasets are always interested in finding out the data distributions across different combinations of dimensions/variables via binning (also called as bucketing).

Traditional techniques such as bar charts and line graphs are often not sufficient for handling multi-dimensional data, which can have numerous variables and large sample sizes. Instead, more advanced visualization methods such as sankey diagrams, parallel coordinates, and heat maps are needed to effectively represent and analyze the data. Even these advanced visualization methods have drawbacks and limitations to what they can represent. The project aims to enhance their capabilites via interactions and custom designs.

The data can be divided into bins in a number of ways using one or more dimensions, and there are many possible ways to form these groups, such as dividing numerical variables into ranges or grouping discrete variables in different ways. However, there is a restriction on the amount of binning that can be performed on the dataset. As the groupings become more and more specific, there may not be enough samples in each group to draw meaningful conclusions or make comparisons. Moreover, as the groupings become more and more general, there is loss of correlation between datasets and hence loss of useful information. Thus, there needs to be a right set of balance on the binning of dataset across different dimensions.

The tool aims at providing a bridge between the binning computations and visual perception of distributions created by those algorithmic computations. The goal of tool is to create visualizations that aim in showing the distribution of multi-dimensional data set across different groupings formed by combinations of various dimensions/variables. The tool involves generating multiple views or perspectives of the data, each of which highlights different aspects or grouping distributions of the data. The views are interactively linked and combined to provide a more comprehensive understanding of the data. It addresses important design decisions, upside and

downsides for different design decisions and potential areas for future development. The focus is on the fundamental principles and development of a tool that aids user in exploratory analysis of data via grouping and at the same time act as the bridge in between computations (binning algorithms) and data visualization.

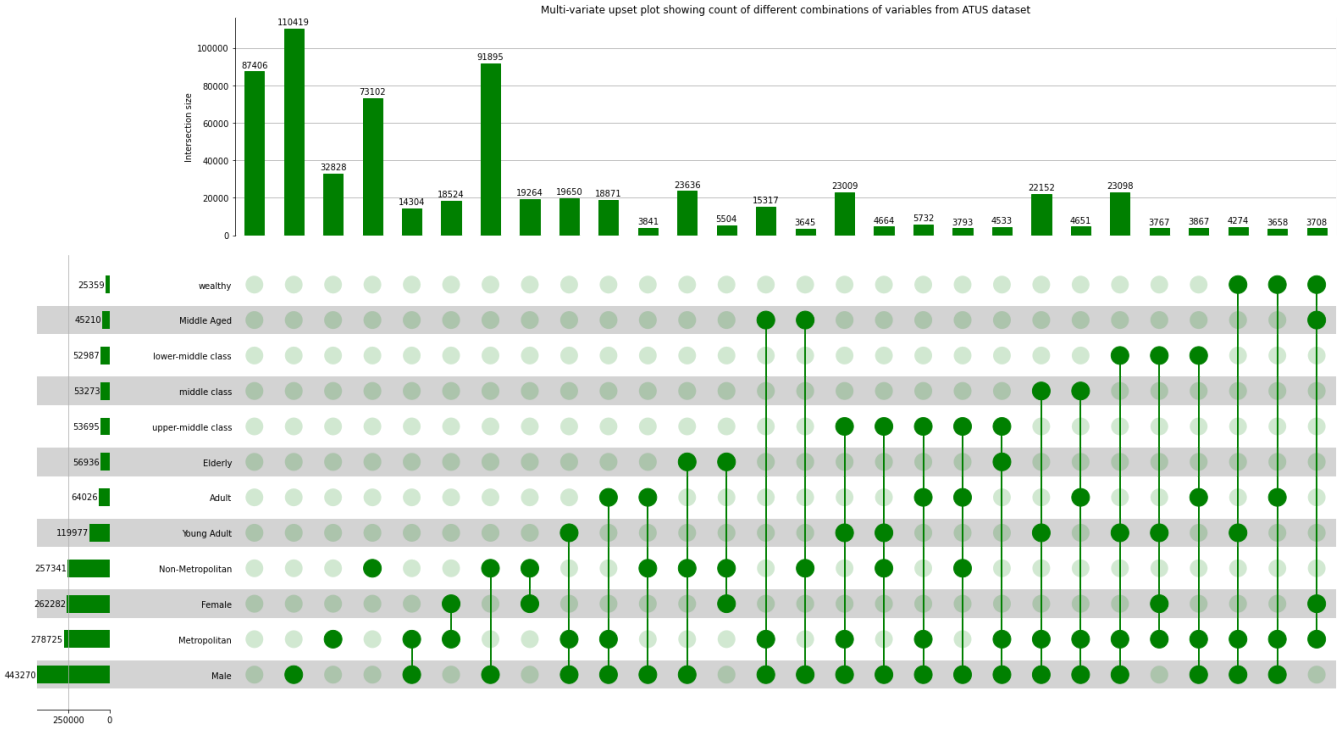

Figure1: Upset Plot for 4D attributes

2 Challenges

The main challenges involved in designing such a toolkit is to present multi-dimensional data onto 2D dimensional space (assuming that the data visualizations represented by the toolkit is 2D). To do this, we present and design different data visualizations that either represents multi-perspective data visualizations (via interaction for the selection of dimensions and bins by user) or via filtering the dataset (by selecting the min and max range for data points in each bin). Moreover, as we add additional dimension across the dataset, the number of possibilities explodes exponentially as for each possible route.In a case where there are N dimensions each with M possibilities, there are N^M possible combinations for different combinations.

3 Binning Toolkit

"clarity, precision and efficiency - Edward Tufte."

Idea: The main idea behind the toolkit is to provide a multi-perspective and interactive approach for exploring and perceiving multi-dimensional data distributions. Binning is fundamental approach in gaining insights about the data distributions and and outlier detection. The number of ways to perform binning explodes exponentially with the increase in the number of dimensions and directly depends upon the bin range per dimension. Things get visually intractable to comprehend when the dimensions are continuous variables and there can be infinite bins possible. To keep the tool simple and build upon the fundamentals of data binning, the tool is designed to work with a set of 5 dimensions of the dataset. These dimensions are pre-defined into different bins to enable tractable computations for binning and to enable ease of understanding for the user. For example, categorization of age dimension into predefined discrete categories is essential for the toolkit to present the data spread over the categories for all unique combinations across different dimensions. Toolkit is developed in python language in Dash framework [1].

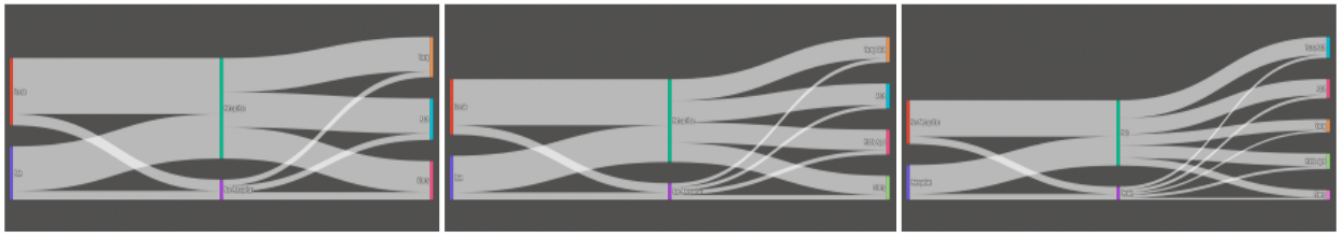

Figure2: Sankey plot with equal width binning (left: 3 bins, center: 4 bins, right 5 bins) on AGE attribute

There are two main components involved in the proposed toolkit - Binning algorthims and data visualizations. Binning algorithms represents the core computing logic behind different strategies employed to group data points across different multi-dimensional bins. The data visualization subpart acts as the bridge in between the core computing logic of binning and the human perception. It is via data visualizations toolkit provides multi-perspective approach for perceiving the multi-dimensional data distributions. Interactions plays a key role here since there are exponential number of different paths in the heirarchical tree for different dimensions. Interactions makes it possible for the user to explore different paths via decisions on which dimension to consider and what kind of binning is performed.

3.1 Binning Algorithms

Pre-defined binning : The tool starts with a given set of dimensions (5 dimensions) and discretizes the dimensions in a predefined set of bins. The ranges for these bins are predecided on the basis of the dimension into consideration. Income levels and age dimensions are discretized into bins based on binning algorithm. Moreover, We add user interaction to control the number of bins and hence control the bin width. The category is computed using equal width binning strategy by equally dividing the range of attribute into equal parts and counting the number of data points within each bin.

The number of bins controls the bin width and thus directly impacts the number of data points per bin. The tool provides interaction to the user to control number of bins to discretize the data attributes. This indirectly allows user to control bin width and thus coverage of data in each bin

$$width_{bin} = (max - min) / num_{bins}$$
$$coverage_i = count(point\ x_i) \quad x_i \in [min + width_{bin} * (i-1),\ min + width_{bin} * i]$$

Custom binning : In this type of binning, the user interacts with the tool to provide a custom [bin_min,bin_max] range and the tool selects a combination of variables which ensures that the data distribution in each bin lies within this range. Internally, aggregations are performed for each bin and bins whose aggregation count lies outside the range are ommitted.

$$filter_{ith\ bin} = true \quad iff.\ coverage_i \notin [user_{min}, user_{max}]$$
$$= false \quad iff.\ coverage_i \in [user_{min}, user_{max}]$$

3.2 Data Visualizations

Data visualizations plays a key role in perceiving the data distributions across multi-dimensions. All of the data visualizations represented by the toolkit are on a 2D dimensional space (no 3D visualizations). This presents challenges on how to convey the data present in 5D dimensional plane onto a 2D dimensional space while preserving the correlation in between data points. The toolkit starts with a representative set of 5 attributes that are chosen to be randomly distributing the dataset. The main objective behind choosing the attributes are to avoid highly correlated attributes (one would want to avoid choosing obvious directly or indirectly correlated values like weekly salary and yearly salary).

Dimension	Data Visualization [hyperlink to viz] (info about viz)	Attributes Presented	Custom Interactions (if any)	Critique
1D	Bar chart [link] *(spread of data in different bins per 1D dimension)*	Single Attribute	which attribute to present in viz	+ relative comparison between bin coverage + exact magnitude of bin coverage - not scalable with increasing bins
2D	Heat Map [link] *(color encoding - color hues to represent data distribution)*	All possible ($^5C_2$ combos) combinations of 2 attributes	which two attributes to present in viz	+ highlight patterns/trends for bin coverage - cannot get exact value of bin coverage - relative comparison between bins is not easy
	Adjacent Bar Chart [link] *(color encoding to represent one dimension)*	All possible ($^5C_2$ combos) combinations of 2 attributes	which two attributes to present in viz	+ relative comparison is easy - not scabale to attributes with multiple bins - comparison difficult between different values of dimensions
3D	Sankey Diagram [link] *(spread of data across 3 dimensions. Interactions to control binning on AGE)*	TESEX, GTMETSTA, AGE *(pre-binned and customizable)*	The number of bins for AGE *(calculated according to equal-width binning strategy)*	+ user interaction to control binning + easy to see the flow of coverage across attributes - can get cluttered with increasing number of bin granularity
	TreeMap [link] *(spread of data across 3 dimensions with focus on part-whole relations on heirarchical data. Interaction to change point of focus)*	TESEX, GTMETSTA, AGE *(pre-binned and customizable)*	Changing the area of focus for treeMap *(zooming into particular heirarchical relation to generate new TreeMap)*	+ good to see heirarchical relationships and their bin coverages + shows part-whole relationship between coverage + interaction allows change of focus of root relation - not scalable to fine bin granularity
4D	Upset Plot [link] *(data spread across different combos of data across 4D bins. Interaction to filter number of samples per bin)*	TESEX, GTMETSTA, AGE *(pre-binned)*, SALARY *(pre-binned)*	Filter for [min,max] range for the bin coverage *(slider for inputting the bin coverage range)*	+ interaction allows filtering on bin coverage range + scalable to 5D and 6D dimensions - difficult to interpret and need statistical knowledge - not very effective to show precise values and trends
	Circular Packing [link] *(data spread presented in heirarchical circular packing. Interaction to change the point of focus)*	TESEX, GTMETSTA, AGE *(pre-binned)*, SALARY *(pre-binned)*	Changing the root level area of focus *(radio button to select the filter on top-level heirarchical attribute)*	+ good to see heirarchical and part-whole relationships and their bin coverages + interaction allows change of focus of root relation - not very scalable to fine granular bins
	TreeMap [link] *(spread of data across 4 dimensions with focus on heirarchical relations. Interaction to change point of focus)*	TESEX, GTMETSTA, AGE *(pre-binned)*, SALARY *(pre-binned)*	Changing the area of focus for treeMap *(zooming into particular heirarchical relation to generate new TreeMap)*	+ good to see heirarchical and part-whole relationships and their bin coverages + interaction allows change of focus of root relation - not scalable to fine bin granularity
5D	Dendogram [link] *(clustering of bins with simlarity on the basis of coverage.)*	TESEX, GTMETSTA, AGE *(pre-binned)*, SALARY *(pre-binned), LEISURE time (pre-binned)*	Not applicable	+ clustering of similar coverage bins would help in equal depth binning - get more cluttered as the number of bins increases

References

[1] Data visualizations using plotly, https://plotly.com/python/plotly-express/
[2] Displaying images using plotly, https://plotly.com/python/renderers/
[3] Plotly framework for interactive python data visualizations https://dash.plotly.com/
[4] Adding interactions into the toolkit using callbacks, https://dash.plotly.com/advanced-callbacks