PART1

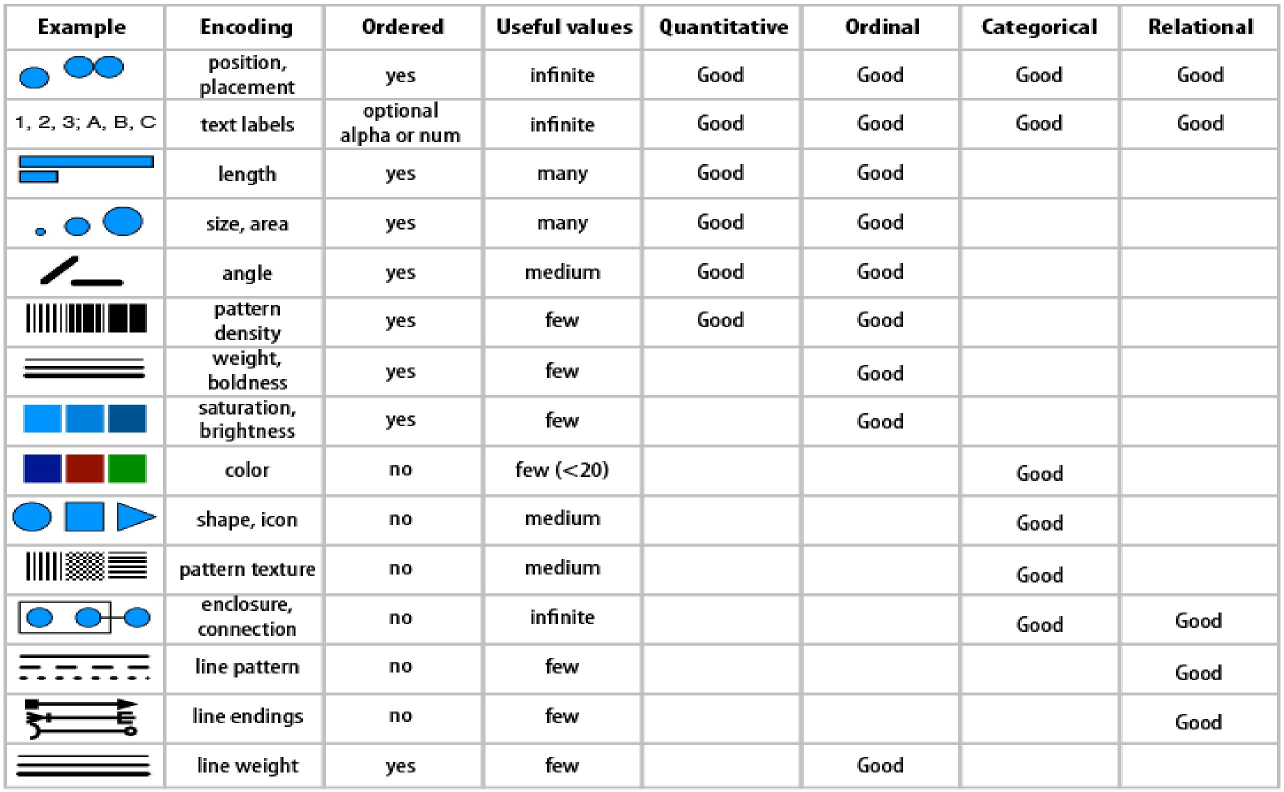
What are the main ways of encoding data in visualization i.e. visual encoding channels?

The eight visual variables are as follows:

POSITION, primarily used in GIS, this gives information on the location of the item.

MARK is a basic graphical element in an image. A 0D mark is a point, 1D mark is a line, 2D is an area, and 3D is a volume. They are helpful to recognize many classes.

SIZE is useful in comparing the intensities of various data points. They represent how large/small a data is drawn.



Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
● ●●	position, placement	yes	infinite	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	optional alpha or num	infinite	Good	Good	Good	Good
▬	length	yes	many	Good	Good		
• ● ⬤	size, area	yes	many	Good	Good		
∕ —	angle	yes	medium	Good	Good		
▥ ■■	pattern density	yes	few	Good	Good		
☰	weight, boldness	yes	few		Good		
🟦 🟦 🟦	saturation, brightness	yes	few		Good		
🟦 🟥 🟩	color	no	few (<20)			Good	
● ■ ▶	shape, icon	no	medium			Good	
▥ ▩ ☰	pattern texture	no	medium			Good	
● ●-●	enclosure, connection	no	infinite			Good	Good
- - -	line pattern	no	few				Good
→	line endings	no	few				Good
☰	line weight	yes	few		Good		

BRIGHTNESS, also known as value, intensity, illuminance, is used in shading representations

COLOR is good for categorization of groups based on colors. This channel works well with qualitative data

ORIENTATION, also called as direction. Mostly used in spatial/geometric data representation

TEXTURE is a combination of other encoding techniques

MOTION variables can be associated with other visual variables for conveying information

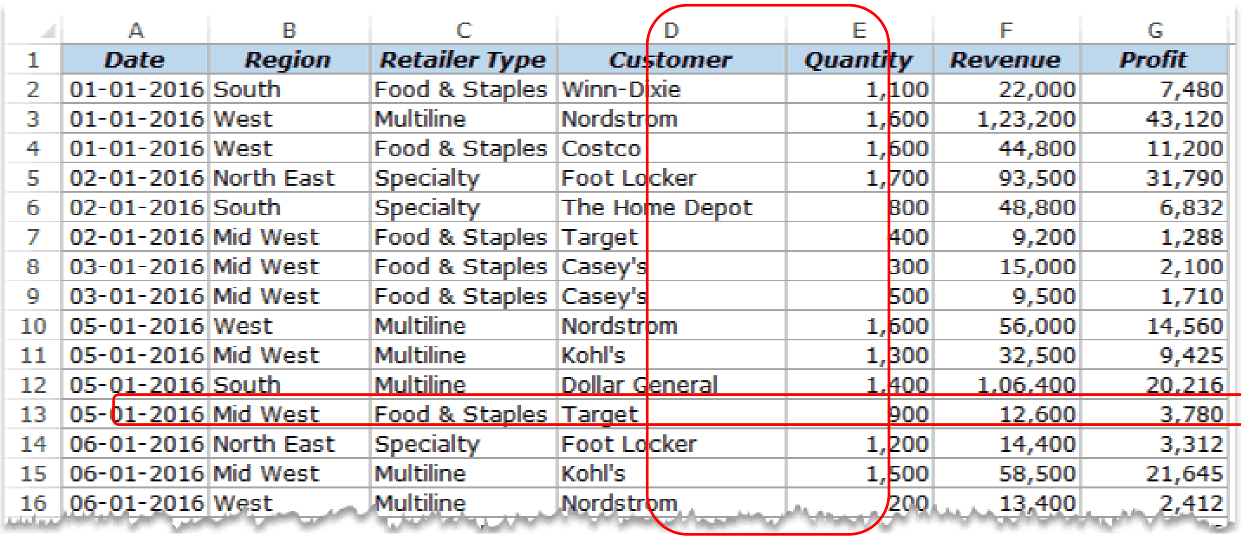
What are some recurring types of datasets in visualization?

There are majorly four basic types of visualization:

Tables

Tables are made up of rows and columns.

While columns are the attributes of the datasets, each row represents a data-point or an item of the data. So, a cell gives the detail of the attribute of that item in a dataset. Tables are flat (1D – with one attribute or with one data point, 2D – multiple attributes and multiple data-points) and multidimensional in nature. Due to range of attributes, it is easier to define the index of the flat tables as compared to the multidimensional ones. Multiple keys are required to look up an item. The combination of multiple keys when unique is used for indexing.



	A	B	C	D	E	F	G
1	***Date***	***Region***	***Retailer Type***	***Customer***	***Quantity***	***Revenue***	***Profit***
2	01-01-2016	South	Food & Staples	Winn-Dixie	1,100	22,000	7,480
3	01-01-2016	West	Multiline	Nordstrom	1,600	1,23,200	43,120
4	01-01-2016	West	Food & Staples	Costco	1,600	44,800	11,200
5	02-01-2016	North East	Specialty	Foot Locker	1,700	93,500	31,790
6	02-01-2016	South	Specialty	The Home Depot	800	48,800	6,832
7	02-01-2016	Mid West	Food & Staples	Target	400	9,200	1,288
8	03-01-2016	Mid West	Food & Staples	Casey's	300	15,000	2,100
9	03-01-2016	Mid West	Food & Staples	Casey's	500	9,500	1,710
10	05-01-2016	West	Multiline	Nordstrom	1,600	56,000	14,560
11	05-01-2016	Mid West	Multiline	Kohl's	1,300	32,500	9,425
12	05-01-2016	South	Multiline	Dollar General	1,400	1,06,400	20,216
13	05-01-2016	Mid West	Food & Staples	Target	900	12,600	3,780
14	06-01-2016	North East	Specialty	Foot Locker	1,200	14,400	3,312
15	06-01-2016	Mid West	Multiline	Kohl's	1,500	58,500	21,645
16	06-01-2016	West	Multiline	Nordstrom	200	13,400	2,412

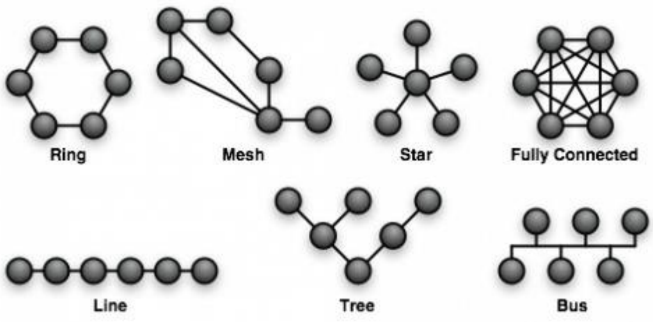
In the above image the vertical mark denotes the attribute of the item, the horizontal box denotes the item and the intersection with the value "Target" is the value of the attribute for that item.

Networks & Trees

Networks: Networks are made up of nodes and links. This type is used to represent the connection between various items via links. An item in a network is called a node. So, links define the connection between nodes. The links are bidirectional, may have attributes associated with them.

Trees: Unidirectional networks are trees. Networks with hierarchical structure are trees. Trees are acyclic; each child node has one parent node pointing to it. Other types of networks are as follows:

Basic Network structure:



Fields:

The field dataset type also contains attribute values of the cells. The cells contain measurements and calculations from a continuous domain. Because of it's continuous nature infinitely many values can be taken, by taking values between two existing ones each time. The more the values the more refined the visualization and analysis be.

In case of data sampling, careful examining is required as continuous data requires strategy for sampling.

Spatial fields:

The cell structure of the field is sampled at spatial positions. The subfield of scientific visualization (aka scivis) is dependent on spatial position in the dataset. The subfield of information visualization (aka infovis) depends on choice of the designer to consider space in visual encoding.

Geometry:

The dataset type specifies the geometry of the of the item using spatial position. The items could be 1D lines or curves, 2D surfaces or regions, or 3D volumes. They require shape understanding. They do not necessarily have attributes. It should be derived or transformed in a way the requires consideration of design choices.

The multivariate structure of the fields depends of the number of value attributes such that a scalar has one attribute per cell, vector has two or more and tensor has many. While the multidimensional structure depends on the number of keys. These include cases such as 2D, 3D fields.

What are the recurring types of tasks in visualization?

Some tasks in visualization are:

Identify: recognize object based on characteristics presented
Locate: establish position of an object in a multidimensional view
Distinguish: determine if object is distinct from another
Categorize: classify objects into different types
Cluster: group objects based on a relationship. The clusters are then treated based on their group properties.
Rank: place a group of objects in an order. This helps in prioritizing analytics type of tasks
Compare: examine the similarities and differences between objects, this can be done using color encoding.
Associate: draw a relationship between objects, primarily done using network links.
Correlate: find a causal or reciprocal relationship between objects

PART 2

Provide two examples of datasets of different types with a corresponding visualization. Discuss the visualization with regard to the concepts of data types, tasks and visual encoding channels as discussed in the module [8 marks]

Provide a short critique for the visualizations: as a minimum you should discuss whether you think the right choice of visual encodings were used and why. You may provide other (subjective) opinions. [2 marks]

I have taken two datasets:

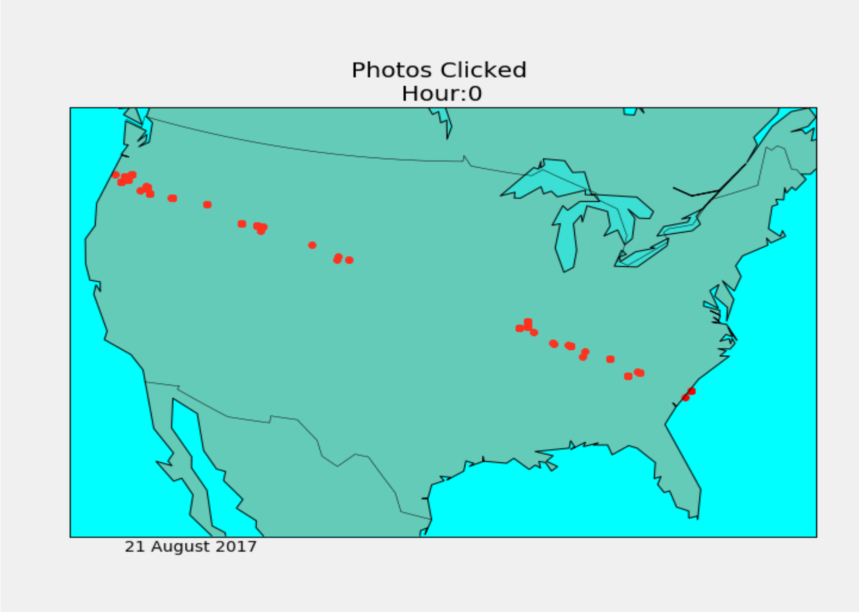
The first one is the Eclipse megamovie dataset (eclipsemega.movie).

I am considering the solution present on Kaggle as in the below link:

https://www.kaggle.com/ash316/kb-mb-gb-tb-b-bigquery/data

The eclipse movie dataset is a collection of photos (~135k), of the total solar eclipse clicked by users in USA from various locations. The number of such users is 1925. The dataset type is spatial in nature.

below figure. The red dots are the location from where the photos have been clicked, they vary at various time intervals. But, the path remains the same, overall.

For visual encoding the position channel has been used. The highlighting of the position by red colored dots on the plane blue map ease locating. In the actual gif from the solution page, another type of encoding can also be seen which is "motion", for determining at what hour what the pictures have been clicked. The visualization gives the required details.	

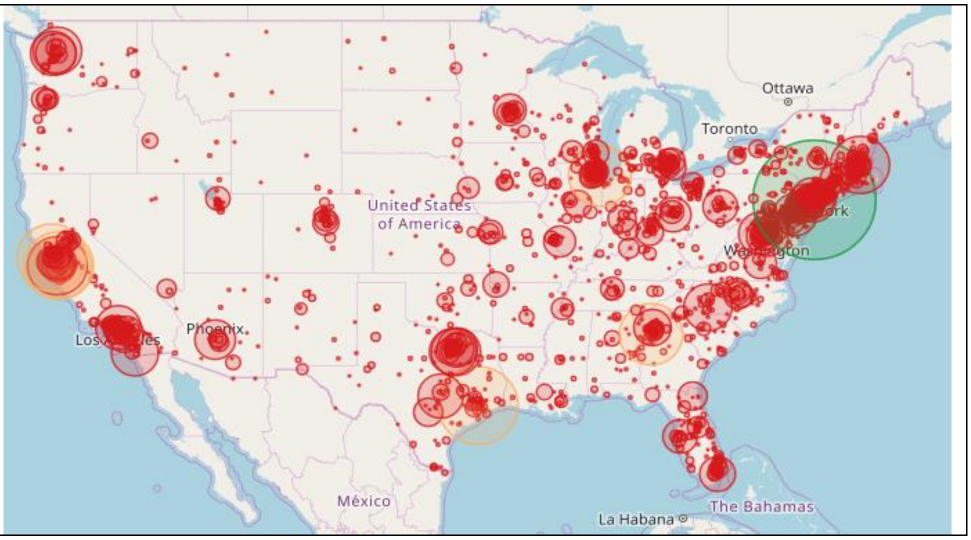
The second dataset is the US data set on H-1B Visa.

https://www.kaggle.com/gpreda/h-1b-visa-applications/data

The dataset is in the form of a table.

I’m considering the below solution.

https://www.kaggle.com/gpreda/h-1b-visa-applications

Two types of visual encodings have been used: Position and Size. Position is for locating the place of the applicants and size is for presenting the quantity of the visa applications from across US. more.	 Ottawa Toronto United States of America York Washington Los Angeles Phoenix México La Habana The Bahamas

The dataset required locating and ranking of the number of the applications. Though, the size of the bubbles signify that the major applicants are from east locations, but major ranking using numbers would have helped