# Selecting the Right Data Visualization in a Data Science Pipeline for Data Exploration and Data Interpretation

## Introduction

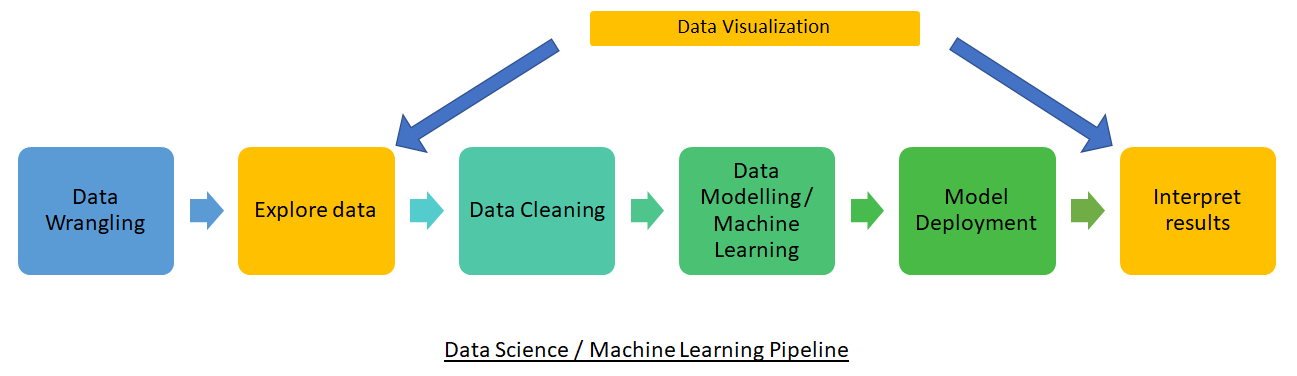
The result of many Data Science projects is to allow good business decision making, and an easily and accurately interpretable Data Visualization is key to that. Choosing the wrong visualization can lead to incorrect business decisions, impacting the company’s bottom line.

As Data Scientists, it is important we understand the science of human cognitive perception before choosing a visualization.

In this article, I will explain the scientific basis and human psychology of why we find certain Data visualizations far easier to read, and then share a laundry list of different visualizations and when should we use them.

## Data Science Pipeline and Data Visualization

Let us look at a typical Data Science project pipeline that consists of Data wrangling, Data exploration, Data Cleaning, Data modeling, Model Deployment, and Result interpretation.



Data Visualization is used twice in the pipeline. Once to understand the source data, so that the correct model can be created. The second time is to interpret the result so that the correct business decisions can be made. If we chose the incorrect visualization in either of these steps, we could end up making an incorrect business decision, and the company could potentially lose a lot of money.

## Quantitative vs Categorical vs Ordinal Variables

The human mind perceives different visualizations differently depending on if the data being plotted is Quantitative, Categorical, or Ordinal. Quantitative data will always be a number that can be measured. Categorical or Nominal data is classified without a natural order or rank, whereas ordinal data has a predetermined or natural order. Examples are given below for ease of understanding

* Quantitative : 1, 2, 3, 4, 5…..
* Categorical: Vegetable, Fruit, Pulses, Grains (note that there is no natural order)
* Ordinal: Small, Medium, Large, X-Large (note the natural order or rank)

## Human Cognitive Perception Accuracy Scale

Based on various experiments as stated in [1], then refined by [2], the human visual cognizance decodes visual information in the following order of accuracy

Table

Description automatically generated

Visual cues that are higher up in the above table are better and allow for a more accurate interpretation of the data being represented.

We make certain observations

1. Position is the best cue to use for all types of variables: quantitative, ordinal, and nominal
2. Length, Angle, Slope, Area, and volume, each represent quantity, hence can be used for quantitative variables. They are at the bottom of the table for ordinal and nominal types of variables and should not be used for nominal, and ordinal data, since they are bad representations of category

## Different Visualizations

Now we will look at a list of different visualizations and talk about which should be used when. The below text and images are from [3]

We will look at visualizations in three different categories

1. Uni relationship visualizations – visualizations used to compare a single variable across different categories. For example, the population across different countries. We look at the following visualizations in this category
   1. Word Cloud
   2. Packing
   3. Cartogram
   4. Choropleth
   5. Pie Chart
2. Bi relationship visualizations – visualizations used to plot 2 different variables against each other. We look at the following visualizations in this category
   1. Bar Chart
   2. Line Chart
   3. Scatter Plot
   4. Table
3. Multi relationship visualizations- visualizations used to plot 3 or more variables against each other. We look at the following visualizations in this category
   1. Stacked Graph
   2. Relative Stacked Graph
   3. Divergent Stacked Graph
   4. Stacked Line Graph
   5. Stream Graph

## Uni relationship Visualizations

### Word Cloud

A word cloud is used when we don’t want to show the relation between two variables, but how a singular variable is packaged with respect to itself

Timeline

Description automatically generated

Image from [3]

A word cloud shows the words that are used more frequently in a larger font. It uses Area as a visual cue, which if we look at our visual perception accuracy table, isn’t at the top.

### Packing

Similar to the word cloud, we can pack data to show relative sizing like below.

Chart, bubble chart

Description automatically generated

Image from [3] – shows the relative population of countries

Packing again uses area as a measure which is low on the perception accuracy scale.

### Cartogram

A Cartogram is a map that is distorted based on data size. Implying that countries with larger data values are represented with larger sizes.

Map

Description automatically generated

Image from [3] – showing relative population size

A cartogram again clearly uses Area as a visual cue.

### Choropleth

A choropleth is a map that changes density/saturation/hue based on data size

Map

Description automatically generated

Image from [3] – showing world population

A Choropleth is a much better representation than a Cartogram since it uses Density/Saturation/Hue instead of Area.

### Pie Chart

A pie chart shows the relative contribution of two or more variables

Chart

Description automatically generated

A Pie chart uses angle which is fairly high up in the perception scale (3rd). 3D pie charts should be avoided since due to foreshortening and perspective distortion they can paint a wrong picture. For example, in the above 2 pie charts, the grey and blue are the same sizes, but the 3D pie chart shows them to be of different sizes.

You can read more about why 3D charts should be avoided in an article that I wrote here: <https://www.linkedin.com/pulse/data-vizualization-why-3d-charts-terrible-idea-praveen-purohit/>

## Bi-relationship Visualizations

### Bar Chart

A bar chart is used when plotting a discrete independent variable on the x-axis against a continuous dependent variable on the y-axis

Chart, bar chart

Description automatically generated

Image from [3]

It uses position and length to indicate data, both of which are at the top of the perceptual accuracy scale. Hence bar chart is a very good visualization to use.

### Line Chart

A line chart is used when plotting two continuous variables against each other, one independent and the other dependent.

Chart, line chart

Description automatically generated

Image from [3]

A line chart benefits from position but not length, unlike a bar chart. Perceptually it tells the brain that we are looking at trends due to the line connection, hence we shouldn’t use a line chart to map discrete variables.

### Scatter Plot

A Scatter plot maps two independent variables that are quantitative. A scatter plot doesn’t plot a function, unlike a line chart.

Chart, scatter chart

Description automatically generated

Image from [3]

The visual cues in a scatter plot rely primarily on position. Density also comes into play when we have clusters.

### Table

A table is a good visualization to use when plotting two discrete independent variables

Table

Description automatically generated with medium confidence

Image from [3]

A table benefits from position only

## Multi Relationship visualizations

### Stacked Graph

A stacked graph is used to map 2 or more quantitative dependent variables on the y axis with an independent variable on the y axis

Chart, bar chart, box and whisker chart

Description automatically generated

Image from [3]

Similar to a Bar Chart, it uses position and length as the visual cues which are at the top of the perception accuracy chart. To differentiate between the two dependent variables, we use Hue as the visual cue.

### Relative Stacked Graph

Is a variation of the stacked bar graph where we emphasize the relative size of the two dependent variables.

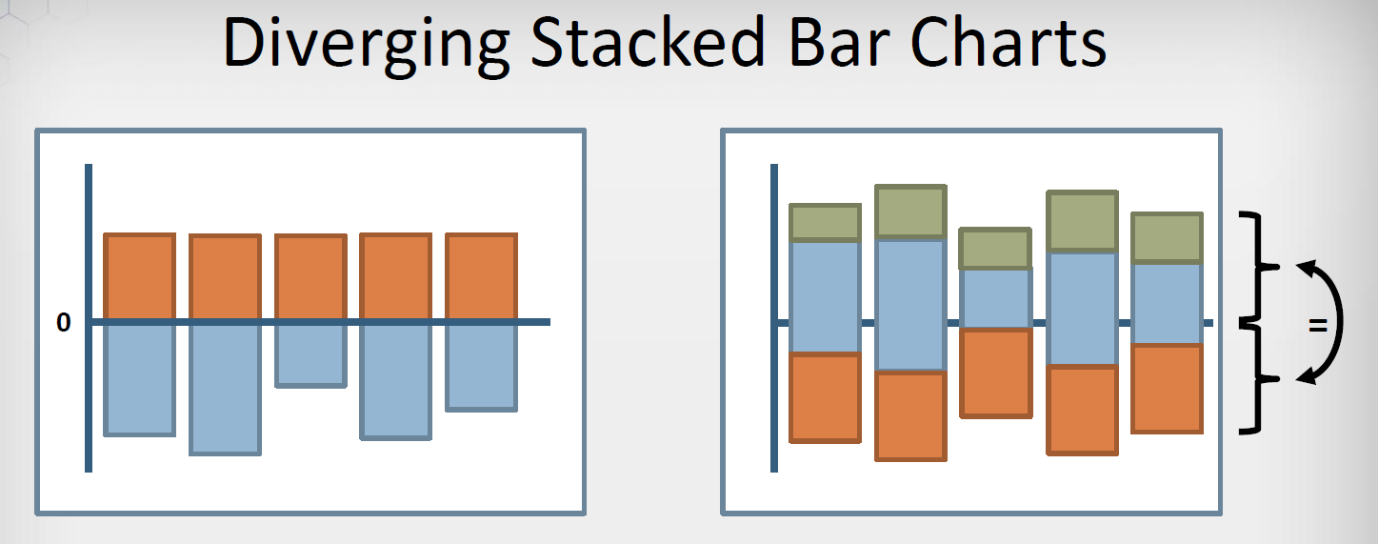
Chart, bar chart, box and whisker chart

Description automatically generated

Image from [3]

### Divergent Stacked Graph

A variation of the Stacked Bar chart is the Divergent stacked bar chart. It gets its name since the bars diverge to both the negative and positive sides of the x-axis.



The bars diverging south of the x-axis have a negative value notation in the left bar chart. The right divergent bar chart allows for easier comparison and is created by first drawing the stacked bars and then centering each bar.

### Stacked Line Chart

If the data is varying continuously, then it is better to use a stacked line chart instead of a stacked bar chart

Shape

Description automatically generated with low confidence

Here it is easier to see the areas as they change, or stay constant.

### Stream Graph

When we have a lot of variables that have continuity and start and finish at different times, it is useful to use a Stream Graph visualization. Stream graph is generally centered to minimize the visual variation in the variables on the outside due to the fluctuation of the variables on the inside.

Diagram

Description automatically generated with low confidence

## Conclusion

In this article, I have captured different visualizations that we can use in a data science pipeline. I have also shared the science behind which visualizations are easier to interpret based on the human perspective scale. I hope you can use this information to make better decisions on which visualizations to use to better tell your data stories, and also ensure that the audience can easily perceive the data.

References

[1] [CLEVELAND, W. S., AND MCGILL, R.](https://www.jstor.org/stable/2288400) Graphical perception: Theory, experimentation and application to the development of graphical methods. Journal of the American Statistical Association, 79(387) 1984.

[2] [J. Mackinlay](https://research.tableau.com/sites/default/files/p110-mackinlay.pdf), Automating the Design of Graphical Presentations of Relational Information, ACM Transactions on Graphics 5(2), 1986.

[3] MCS-Data Science class on Data Visualization by Prof. John C. Hart from University of Illinois Urbana Champaign.