Chapter 2: Visualization Best Practices

Above all else, show the data–
Edward R. Tufte

Learning Objectives

- Understand the importance of data visualization.
- Learn about various types of data and measurement scales.
- Understand how to encode different types of data in charts.
- Learn about effectiveness of visual encodings.
- Understand Edward Tufte’s heuristic graphical design principles.

Importance of Data Visualization

- Humans are better at processing visual information than in the form of tables, numbers, and text.
- Using pictorial representations such as graphs and charts, organizations benefit by transforming their data into an easy-to-understand form, analysed by a wider audience base.
- It helps organizations understand complex data and identify hidden patterns, which leads to improved decision-making.
- Data visualizations can be misleading, especially when dealing with large volumes of data.

Importance of Data Visualization

- By 2025, around 6 billion or 75% of the world’s population will be interacting with data every day, which is expected to create over 90 ZB of data.[1]
- As per Cisco, by 2022, there would be nearly 5 ZB of Internet Protocol (IP) traffic per year [Cooney, 2018].

[1] ***Source:*** https://www.seagate.com/in/en/our-story/data-age-2025/

Importance of Data Visualization

- According to DOMO's[2] "Data Never Sleeps 8.0" report [Anon, 2020], every digital minute:
  1. Facebook users upload 147,000 photos.
  2. 6,659 packages are shipped by Amazon.
  3. 69,444 users apply for jobs on LinkedIn.
  4. Microsoft Teams connect 52,083 users.
  5. Twitter gains 319 new users.
  6. 500 hours of video is uploaded by YouTube users.
  7. 479,452 people engage with content on Reddit.

[2] ***Source:*** DOMO is a software company that specializes in business intelligence and data visualization.

Dataset

- The data considered for charts in this chapter are from the Indian general elections (Lok Sabha elections) published by Election Com- mission of India.[3]
- Election results data for all general elections of India between 1999 and 2019 have been individually collected to form a combined dataset [Elections_dataset.csv].
- We have also selected six attributes for our analysis.

[3] ***Source:*** https://eci.gov.in/

Lok Sabha elections’ dataset

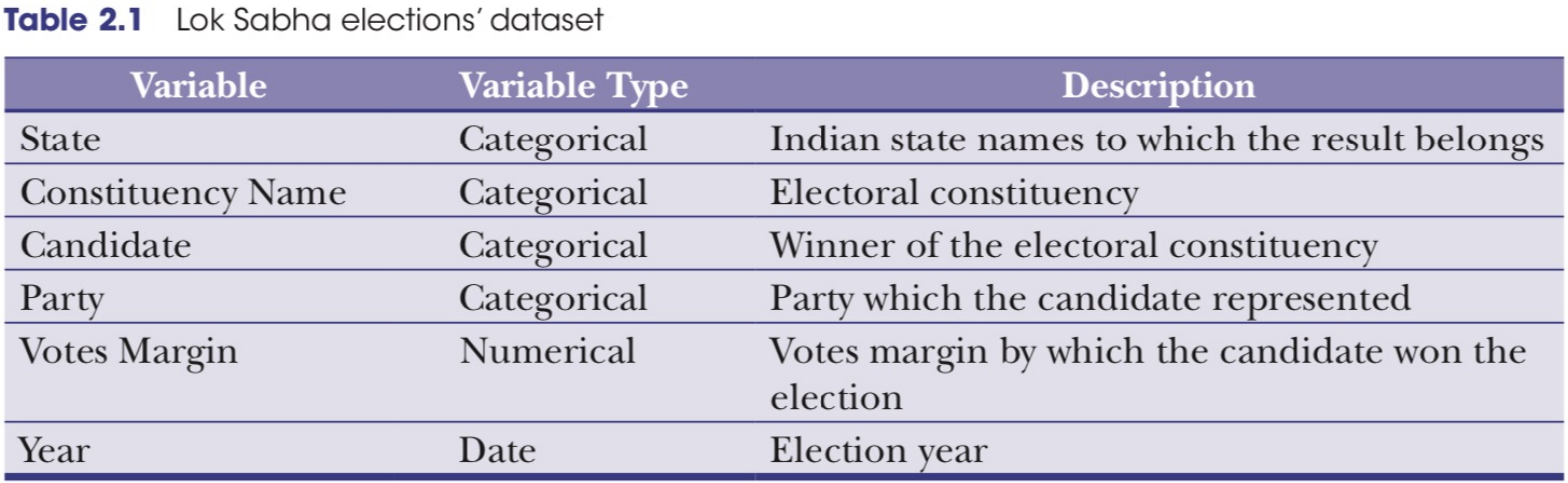


Table 2.1 Lok Sabha elections’ dataset

Variable	Variable Type	Description
State	Categorical	Indian state names to which the result belongs
Constituency Name	Categorical	Electoral constituency
Candidate	Categorical	Winner of the electoral constituency
Party	Categorical	Party which the candidate represented
Votes Margin	Numerical	Votes margin by which the candidate won the election
Year	Date	Election year

Data Types

- **Nominal scale (qualitative variable or categorical variable):** Variable which is qualitative in nature is classified as nominal scale.
  - For example, blood group type which is categorized as A-positive, A-Negative, B-Positive, O-positive, etc is an example of nominal scale variable.
  - We cannot do arithmetic operations such as addition, subtraction, multiplication, or division on a nominal scale variable.
- **Ordinal scale:** These are variables in which the value of the data is captured from an ordered set.
  - For example, review ratings captured on a Likert scale of 1–5 (5 being highest and 1 being lowest).
  - The order of the set is fixed, and we cannot perform arithmetic operations on these variables.
  - As we know, a rating of 5 is better than 3, but a rating of 3 and 2 together cannot be equalled to a rating of 5.

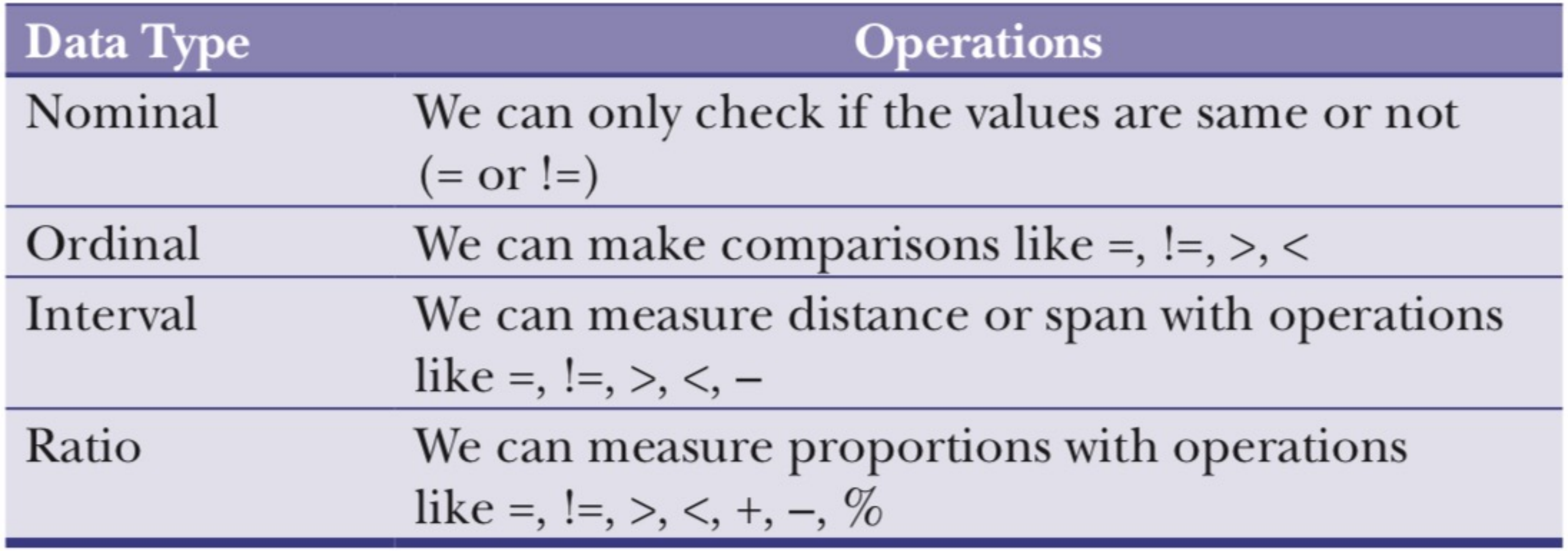
Data Types

- **Interval scale:** These are data points taken from a fixed interval set, for example, IQ level, temperature (in Centigrade).
  - We cannot perform ratios on interval scale data.
  - For example, 50°C is not twice as hot as 25°C. However, 50°C is 25°C more than 25°C.
- **Ratio scale:** Variables for which all four arithmetic operations such as addition, subtraction, division, and multiplication are meaningful are ratio scale variables.
  - A few examples of ratio scale variables are product sales value, employee salary, and market share.

Data Types

- Table 2.2 summarizes details of data measurement scales and corresponding operations which we can perform on these data types.

Table 2.2 Data types and operations



Data Type	Operations
Nominal	We can only check if the values are same or not (= or !=)
Ordinal	We can make comparisons like =, !=, >, <
Interval	We can measure distance or span with operations like =, !=, >, <, –
Ratio	We can measure proportions with operations like =, !=, >, <, +, –, %

Effectiveness of Visual Encodings

- To create effective visualizations, first we need to understand how humans decode graphs.
- William S Cleveland in his book “The Elements of Graphing Data” [Cleveland, 1994] defines three visual operations of pattern perception through which humans decode a graph.
  1. Detection
  2. Assembly
  3. Estimation

Detection

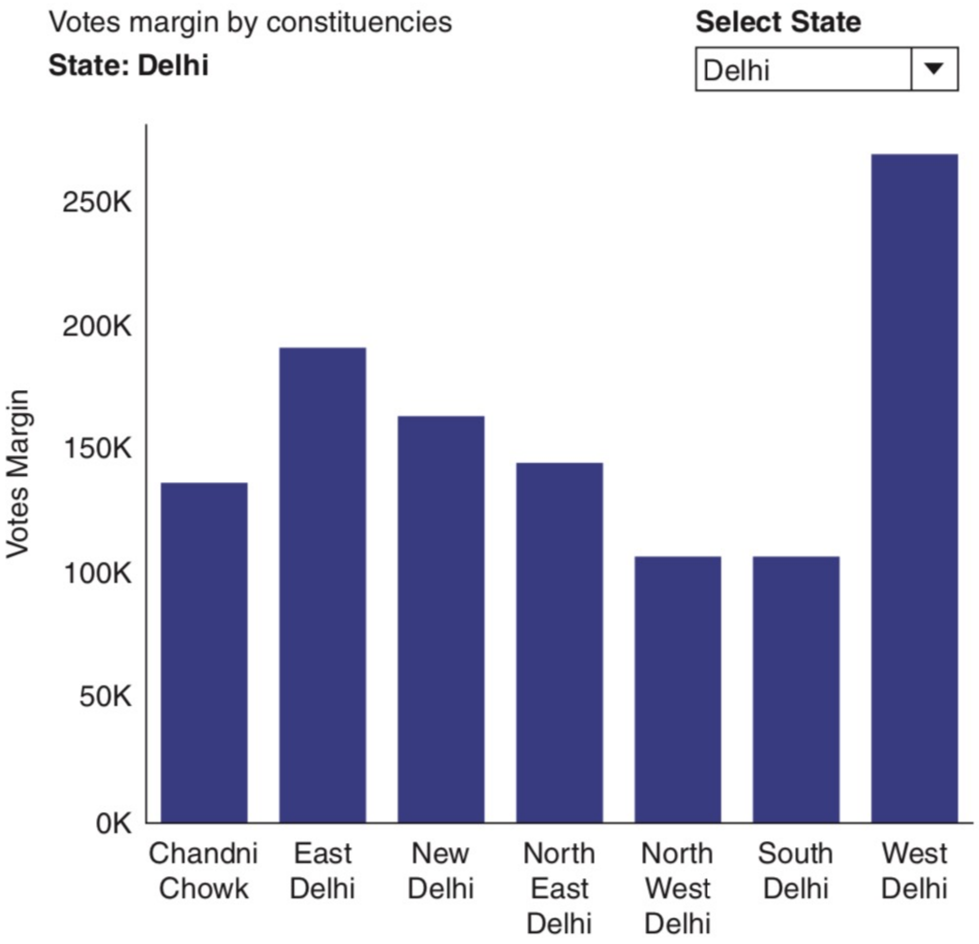


Figure 2.1 Winners' votes margin across all constituencies of Delhi for General Election 2014.

- Detection is about visual recognition of the geometric object encoding the physical values.
- Figure 2.1 shows the chart where we represent vote margin of winners across each constituency of the state of “Delhi” for the election year “2014.”
- Notice that the length of each bar represents the vote margin data across different constituencies of “Delhi.”
- This is also an interactive chart which will provide “state” selection as a dropdown

Assembly

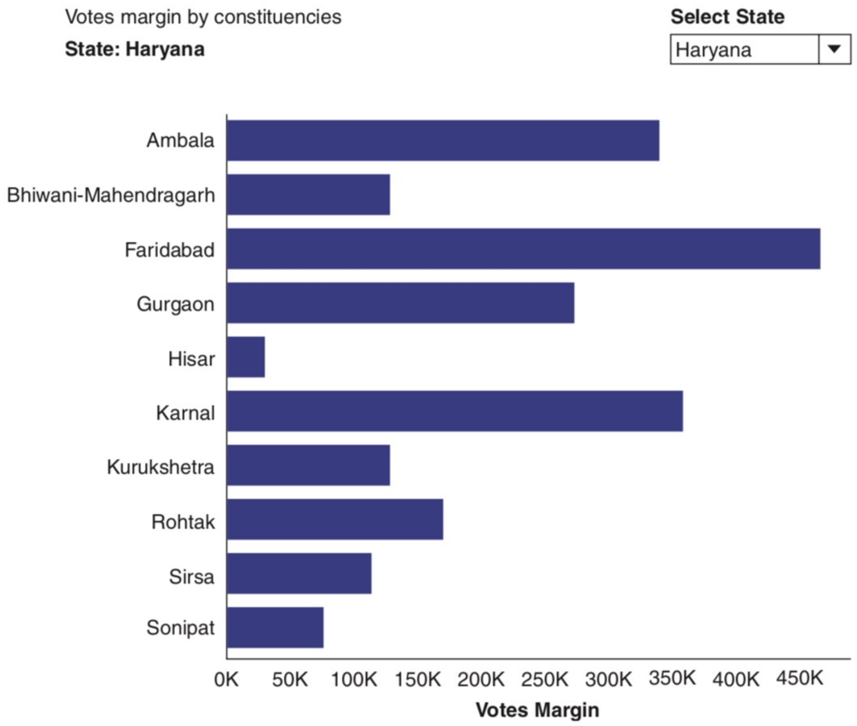


Figure 2.2 Winners' votes margin across all constituencies of Haryana for General Election 2014.

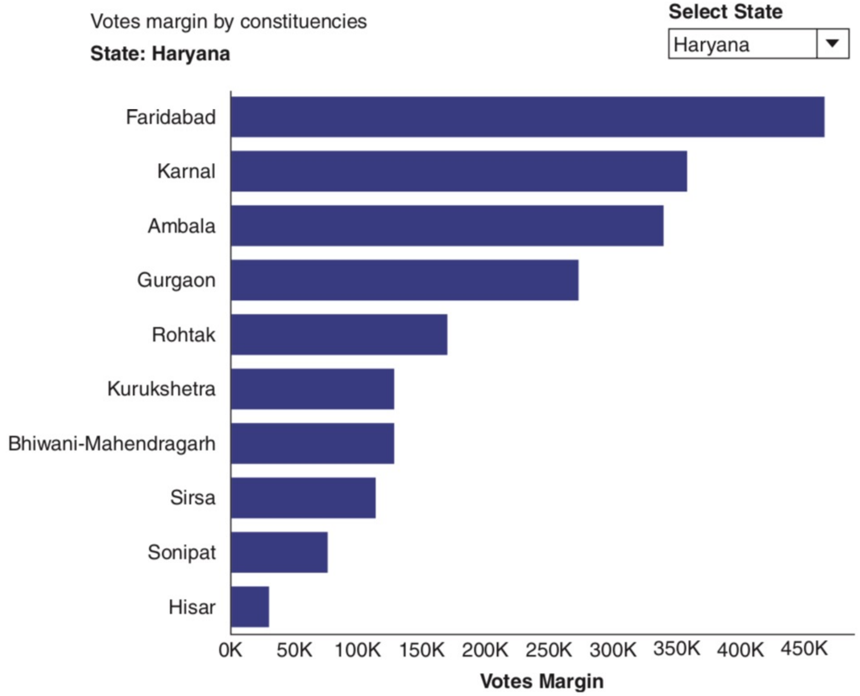


Figure 2.3 Winners' votes margin across all constituencies of Haryana for General Election 2014 – Sorted on votes margin.

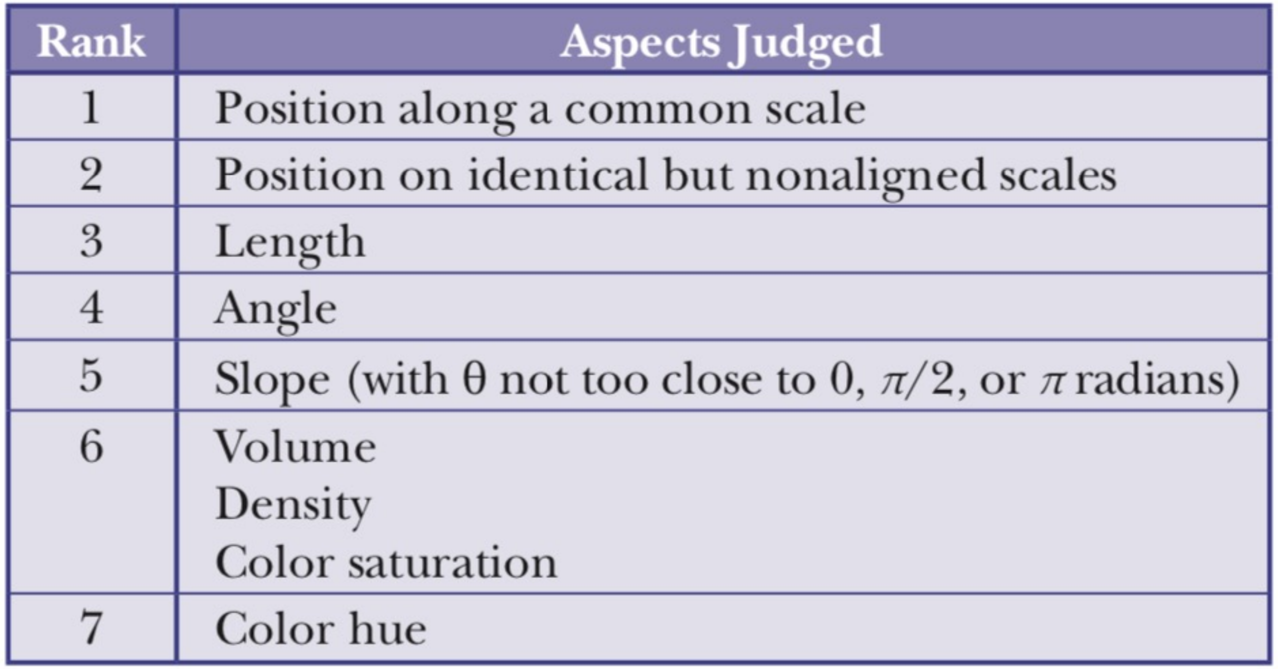
Assembly

- Grouping of detected graphical elements is assembly.
- It is about how we visually simplify charts which will aid our audience in predicting the underlying relationship.
- Let us consider the charts in Figs. 2.2 and 2.3. Both represent the same data, the Votes margin of all constituencies of “Haryana” state in the 2014 elections.
- As per the law of continuity, humans group together those that follow an established direction.
- Hence, in Fig. 2.3, sorting data by vote margin makes it effortless for our audience to understand the underlying data as compared to Fig. 2.2.
- Good plots leverage the law of continuity to assist the assembly.

Estimation

- Estimation is about visually assessing the relative magnitude of two or more quantitative values.
- When we think with data, we are always making comparisons.
- A visualization must feature a comparison between different data points.
- There are multiple levels at which we do comparisons:
  1. Discrimination: $A = B$ or $A \neq B$
  2. Ranking: $A > B$ or $A < B$
  3. Ratio: A/B
- When creating graphs, we should first think what the most important comparison is that we want the user to make and encode that as position on a common scale

Estimation



Rank	Aspects Judged
1	Position along a common scale
2	Position on identical but nonaligned scales
3	Length
4	Angle
5	Slope (with θ not too close to 0, $\pi/2$, or π radians)
6	Volume Density Color saturation
7	Color hue

Figure 2.4 Ranking of data encodings.

Source: Table representation as per Cleveland and McGill (1985) – Cleveland, William and McGill, Ron. (1985). *Graphical Perception and Graphical Methods for Analyzing Scientific Data.* Science (New York, N.Y.). 229. 828–33. 10.1126/science.229.4716.828.

- William S. Cleveland [Cleveland, 1985] lists seven different ways to encode the data and contains the ranks for the encoding from more accurate to least accurate when used for estimation. (Fig. 2.4)

Example Dataset Description

- In our example dataset [Elections_dataset.csv], we have the following information:
  1. State – Nominal
  2. Constituency – Nominal
  3. Candidate – Nominal
  4. Party – Nominal
  5. Votes margin – Ratio (Quantitative)
  6. Year – Interval (Quantitative)

Goal/Question: Analyse and uncover patterns in votes margin data across different states, constituencies and across different years of election.

Color

- Let us plot a chart of votes margin for each constituency by state.
- We will use Color (diverging Color scale) to encode votes margin.
- We have selected state as “West Bengal” and election year as “2014” for our analysis.
- Figure 2.5 is the chart we get when we select “West Bengal” as the state.
- Let us see if we can extract some value out of this graph by applying Cleveland’s estimation tasks.

Color

Discrimination: Let us try and discriminate between two constituencies.

- Is there any difference between Arambagh and Asansol?
  Yes.
- Now, let us try discrimination between Bangaon and Bishnupur.
- In this case, the colour encoding makes the comparison very difficult.

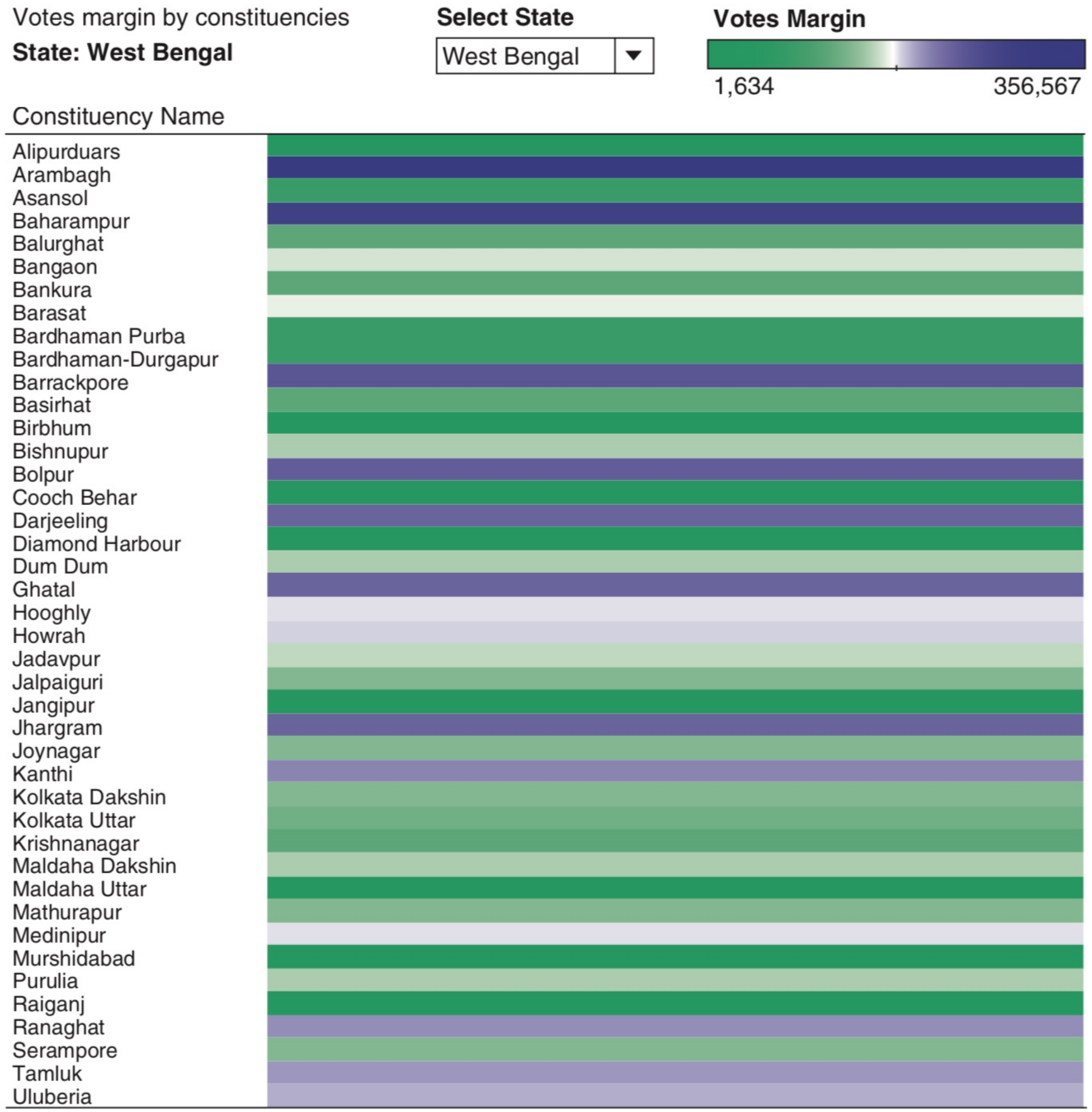


Figure 2.5 Winners' votes margin across all constituencies of West Bengal for General Election 2014 – Votes margin is encoded using diverging color scale.

Color

Ranking: Now, try ranking Dum Dum v/s Diamond Harbour.

  - Which has more votes margin? Yes, it is Dum Dum.
  - You can do it accurately, but it is slow as it makes you refer to the legend to make sure which colour hue is higher or lower.
  - This is true for any encoding with colour hue as hue will not have natural ordering.

- The issue is that the vertical axis labels are sorted in alphabetical order, which is the default encoding for any categorical variable.

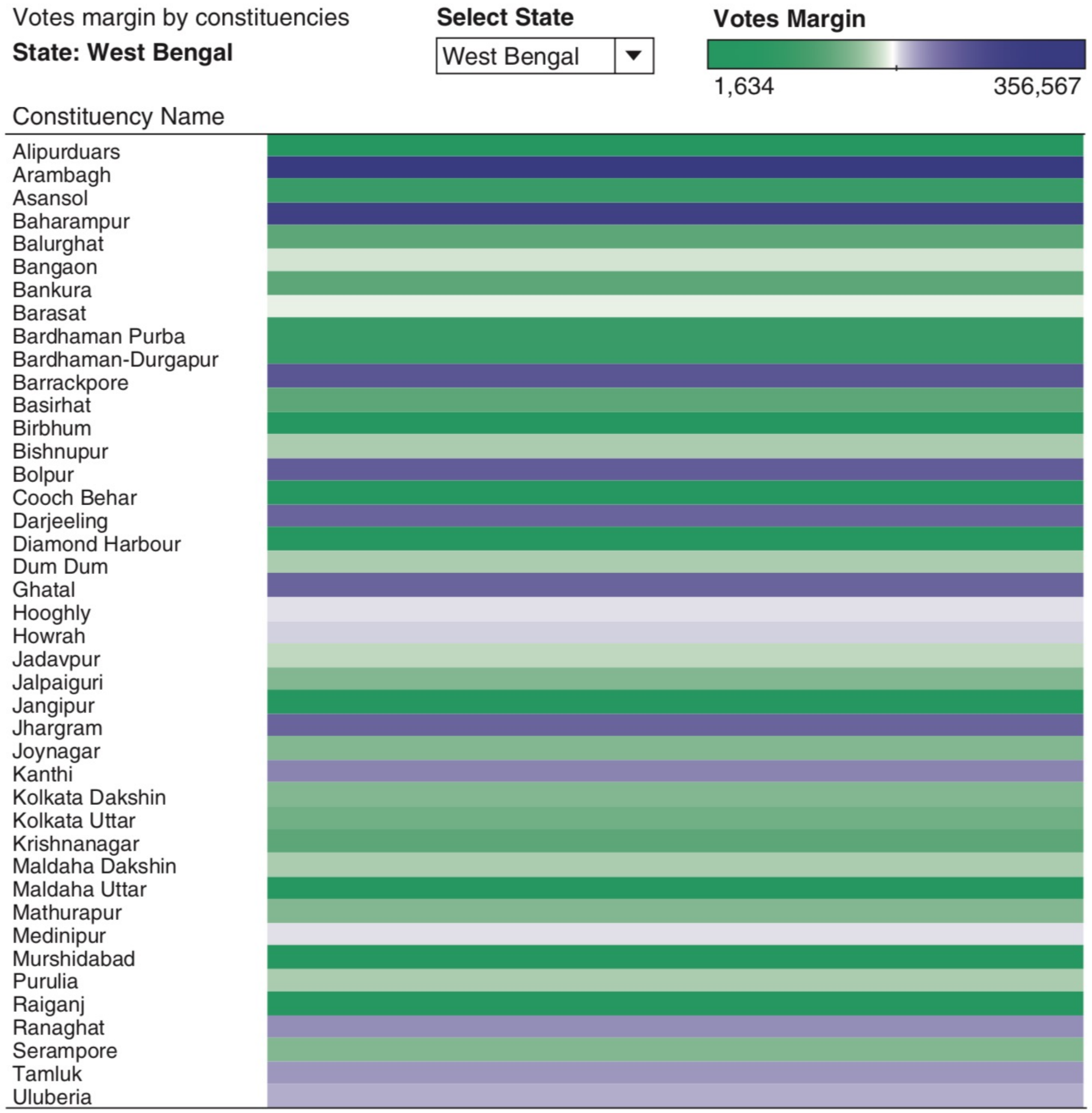


Figure 2.5 Winners' votes margin across all constituencies of West Bengal for General Election 2014 – Votes margin is encoded using diverging color scale.

Color

- When we sort the data on votes margin, we get the plot as shown in Fig. 2.6.
- Ranking becomes trivial if we order/sort on the measure variable votes margin.
- But even after sorting, discrimination is very hard with this plot.
- Hence, we can see that colour is not the best encoding when we need our users to view our visualization and understand measure through comparison.

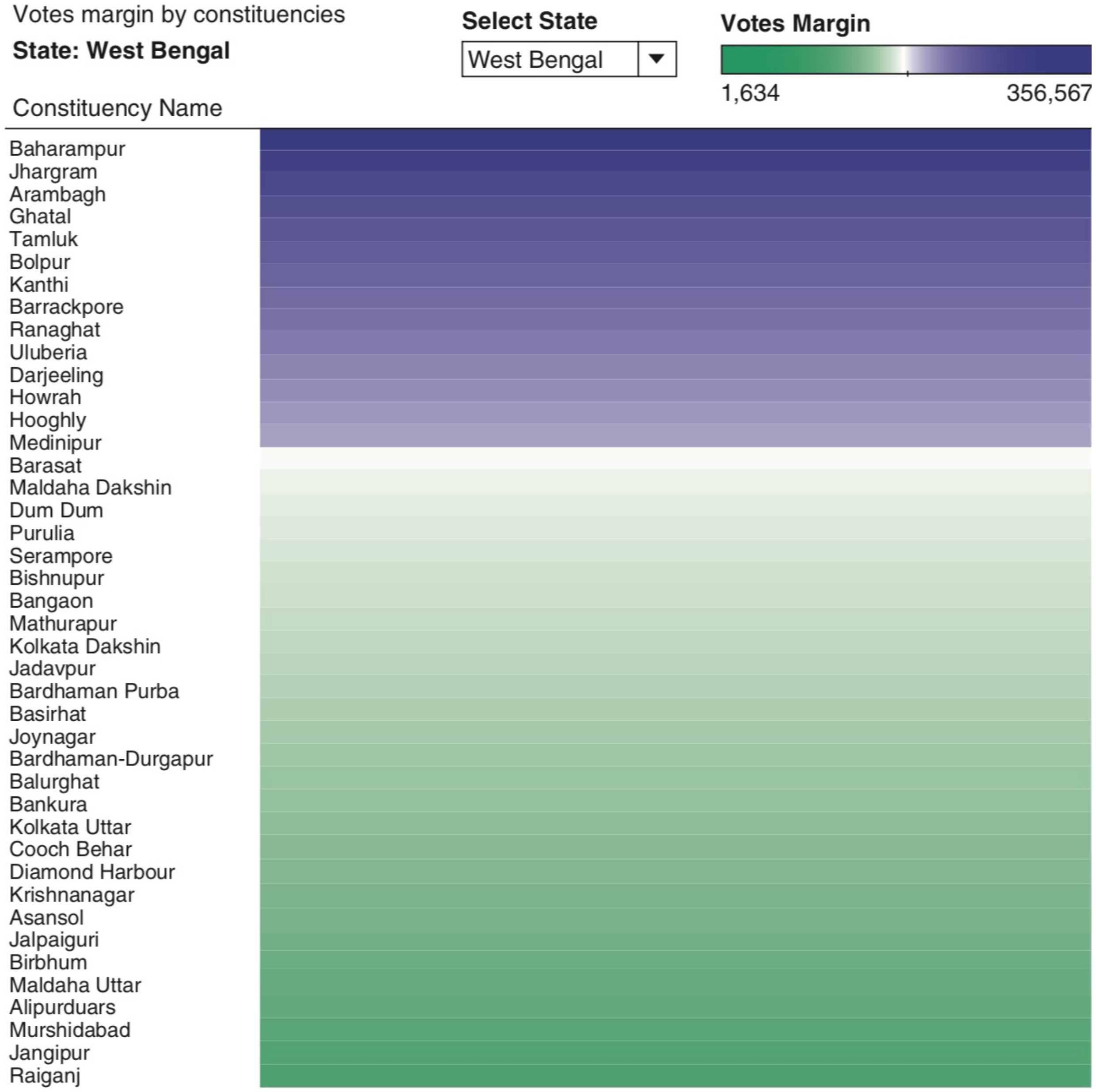
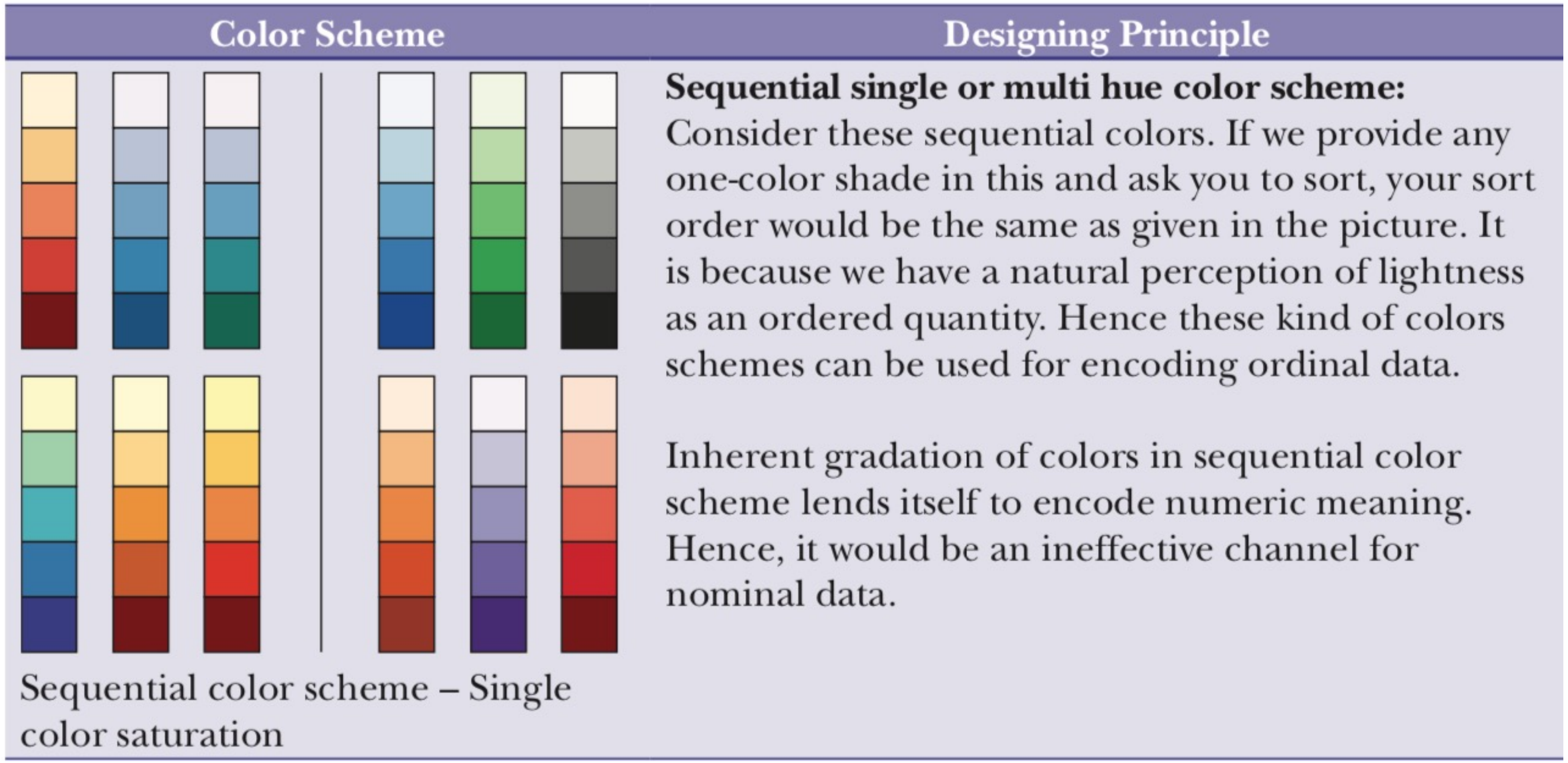


Figure 2.6 Winners' votes margin across all constituencies of West Bengal for General Election 2014 – Sorted by votes margin.

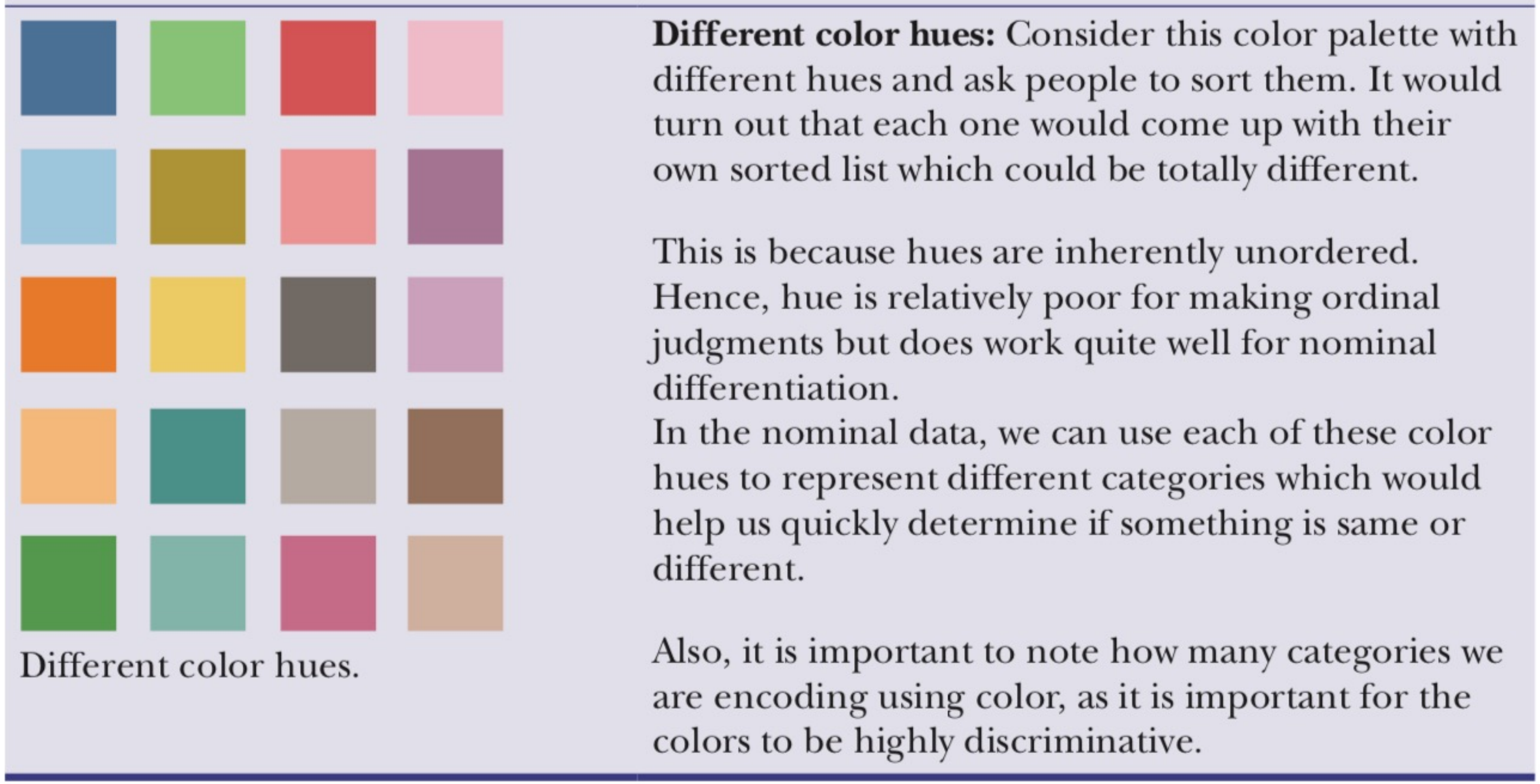
Color Design Principles

Table 2.3 Color design principles



Color Scheme	Designing Principle
Sequential color scheme – Single color saturation	**Sequential single or multi hue color scheme:** Consider these sequential colors. If we provide any one-color shade in this and ask you to sort, your sort order would be the same as given in the picture. It is because we have a natural perception of lightness as an ordered quantity. Hence these kind of colors schemes can be used for encoding ordinal data. Inherent gradation of colors in sequential color scheme lends itself to encode numeric meaning. Hence, it would be an ineffective channel for nominal data.

Color Design Principles



Different color hues.

Different color hues: Consider this color palette with different hues and ask people to sort them. It would turn out that each one would come up with their own sorted list which could be totally different.

This is because hues are inherently unordered. Hence, hue is relatively poor for making ordinal judgments but does work quite well for nominal differentiation.
In the nominal data, we can use each of these color hues to represent different categories which would help us quickly determine if something is same or different.

Also, it is important to note how many categories we are encoding using color, as it is important for the colors to be highly discriminative.

Color Design Principles

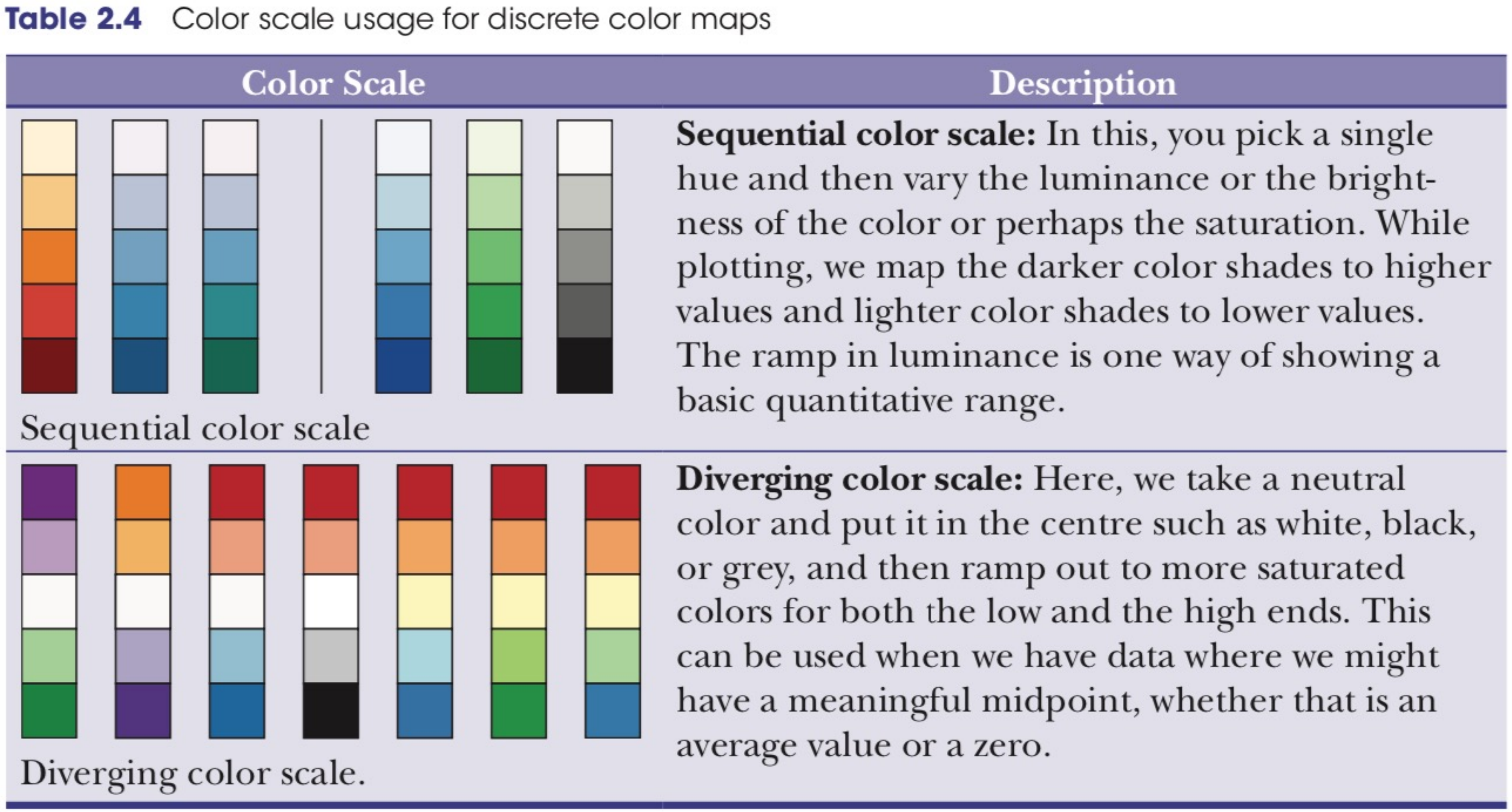
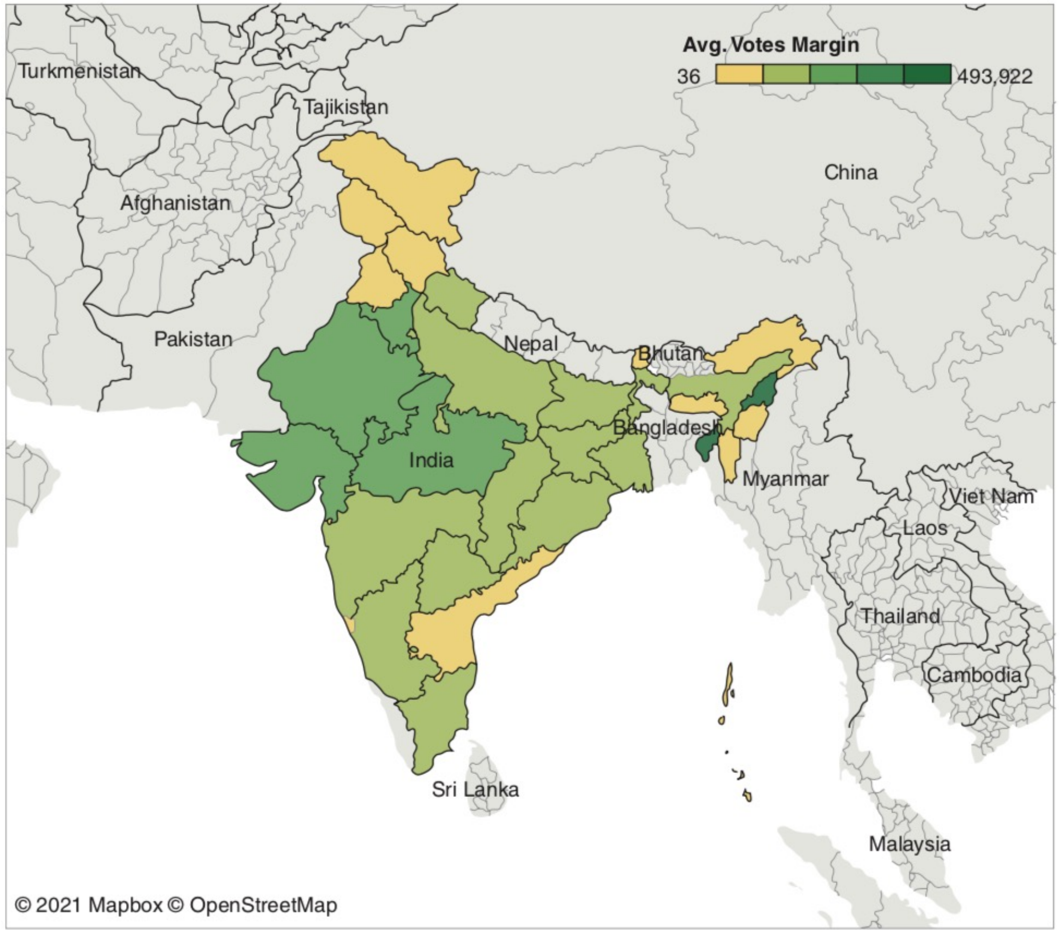


Table 2.4 Color scale usage for discrete color maps

Color Scale	Description
Sequential color scale	**Sequential color scale:** In this, you pick a single hue and then vary the luminance or the brightness of the color or perhaps the saturation. While plotting, we map the darker color shades to higher values and lighter color shades to lower values. The ramp in luminance is one way of showing a basic quantitative range.
Diverging color scale.	**Diverging color scale:** Here, we take a neutral color and put it in the centre such as white, black, or grey, and then ramp out to more saturated colors for both the low and the high ends. This can be used when we have data where we might have a meaningful midpoint, whether that is an average value or a zero.

Color Design Principles

Lok Sabha 2014 - Average votes margin across all states



© 2021 Mapbox © OpenStreetMap

Figure 2.7 Average votes margin for 2014 Lok Sabha election across India.

- In the Figure, average votes margin (quantitative data) for the 2014 Lok Sabha election has been binned into different colour buckets.
- We have grouped the colours into bins and avoiding continuous colour ramps.
- By doing this especially for maps, we are able to keep the colours distinct and therefore help support more accurate comparisons.

Color

- There are a lot of online tools which would help us form the colour palettes like:

1. **Color brewer4:** This helps us contrast different colour schemes and check the robustness of different colour schemes.
2. **Kuler5:** An adobe tools aids in generating colour palette.
3. **Degraveve's colour palette generator6:** This tool helps us to generate colour palette matching an image.

[4] ***Source:*** Color Brewer – https://colorbrewer2.org/#type=sequen- tial&scheme=BuGn&n=3

[5] ***Source:*** Kuler - https://color.adobe.com/create/color-wheel/

[6] ***Source:*** Dagraeve's color palette generator - https://www.degraeve.com/ color-palette/

Color

- Rules of thumb to remember for Color and its perception :
  1. Use a few colours (ideally less than six) and they should be distinctive.
  2. Strive for colour harmony. As a designer, take in some notion of both the data and your own aesthetic sense.
  3. Use cultural conventions. Some concepts may be strongly associated with certain colours. Even if it is not necessarily perceptually ideal, you may still use those mappings, because they are more easily interpreted because of familiarity.
  4. Get it right with **luminance** choices to enable for the colour blind. Even though they might not see the hue, separation in terms of luminance will provide a way for them to distinguish colours.

Volume

- Let us try plotting the same graph as Fig. 2.5 for another state Haryana;
- This time, we use volume to encode the votes margin.
- **Discrimination:** Faridabad has more votes margin than Gurgaon.
- **Ranking:** Faridabad has highest votes margin compared to Hisar, which is the lowest.
- **Ratio:** We can even start without cross-referring to the votes margin legend. For example, Karnal seems to have roughly four times more votes margin compared to Kurukshetra.

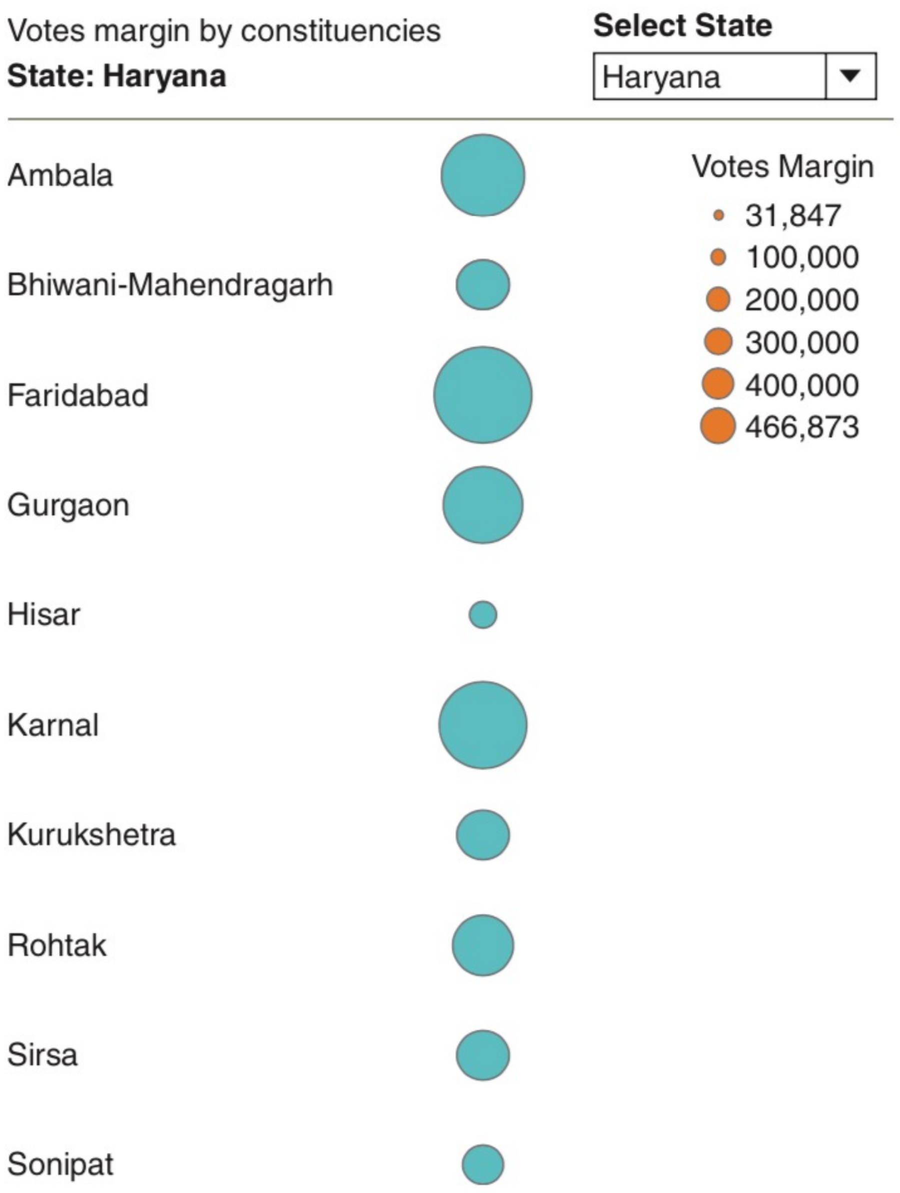


Figure 2.8 Encoding votes margin data as volume.

Volume

- But comparing close areas like Ambala v/s Karnal is still an issue.
- It is extremely rare that we make a plot like this, but two-dimensional area is a very important encoding which can be effectively used to encode information in charts like bubble chart.
- The area of the bubble encodes votes margin (Fig. 2.9).

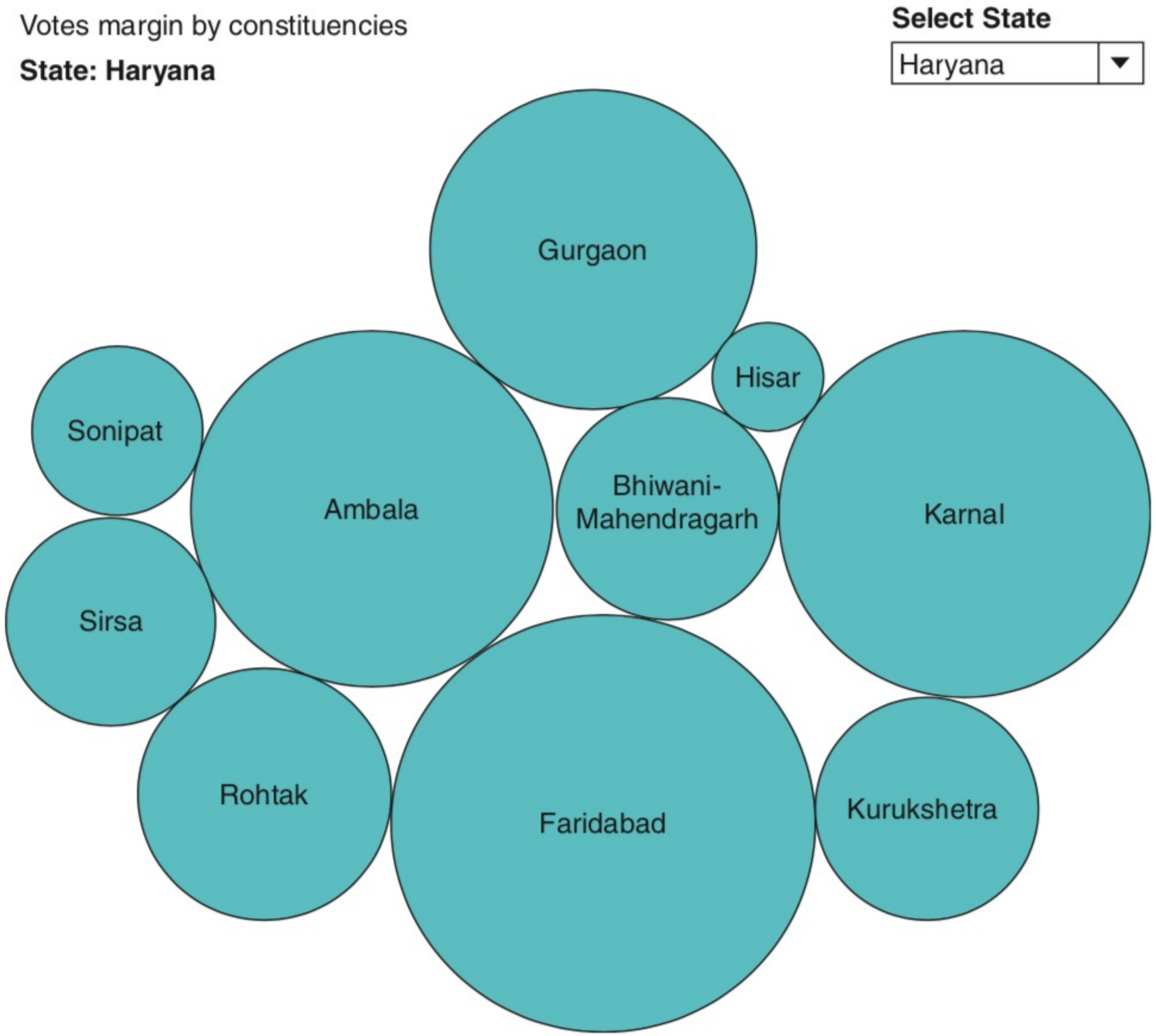


Figure 2.9 Winners' votes margin across all constituencies of Haryana for 2014 elections – Using volume encoding.

Angle

- If we encode the votes margin across years 2009 to 2019 for all "Delhi" constituencies using line chart.
- The votes margin is plotted directly on vertical axis, and hence the rate of change of votes margin is encoded as the angle of the line (Refer Fig. 2.10).
- To understand, which constituency has a better rate of increase in votes margin between Chandni Chowk and New Delhi?
- It is difficult to assess the answer for this question.
- Hence, rather than encoding "rate of change" as an angle, if we plot rate of change in votes margin directly as position on the common scale (Refer Fig. 2.11), the comparison becomes straightforward.

Length

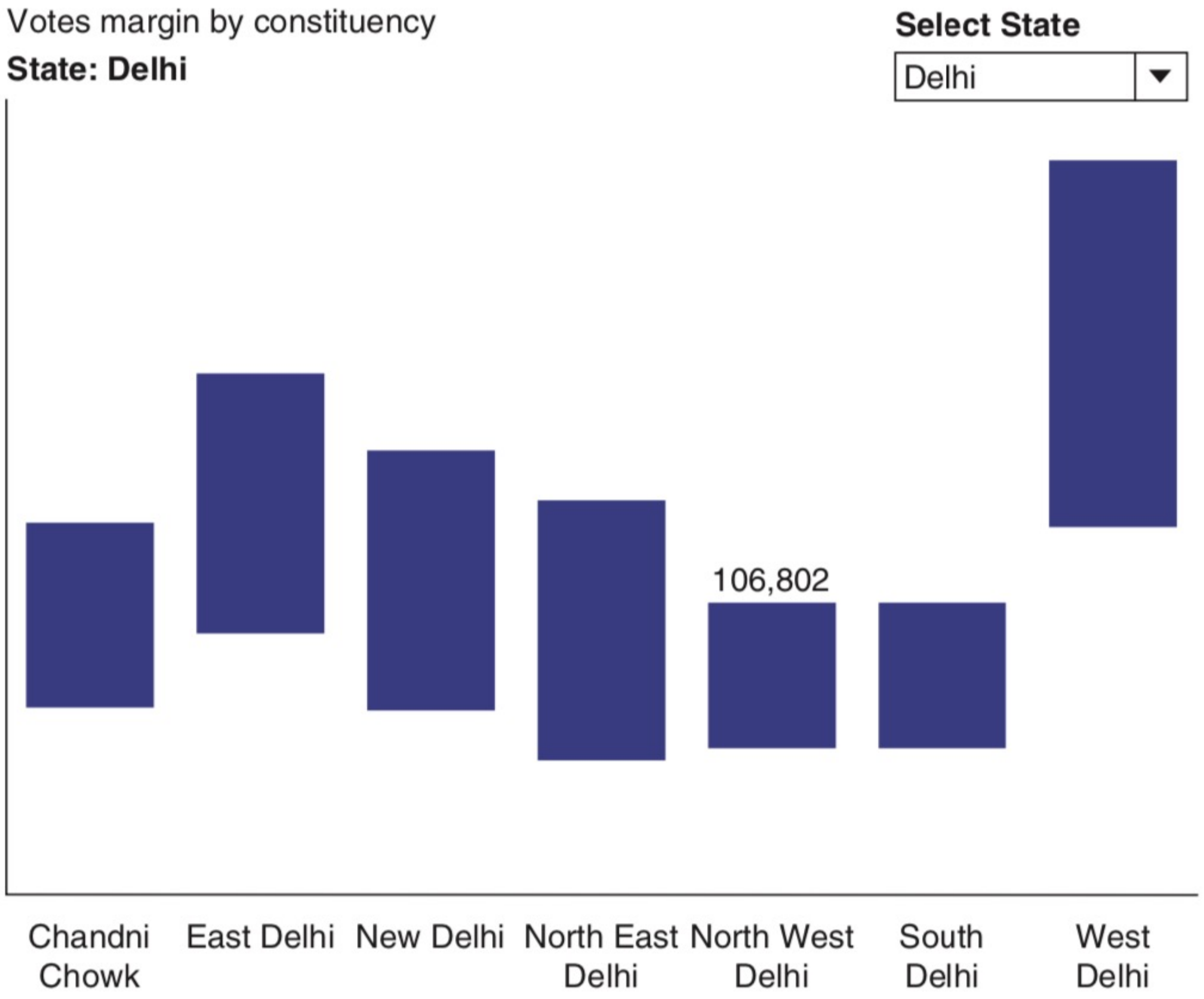


Figure 2.12 Winners' votes margin for all constituencies of "Delhi" for 2014 elections using length as encoding.

- In the chart shown in Fig. 2.12, votes margin has been encoded as the length of bar.
- The bars are jittered and are not against the common baseline.
- The minimum value of the bar is specified, which will help in calculating ratio.
- We can notice that it is much easier to ratio with length compared to area or angle.

Position

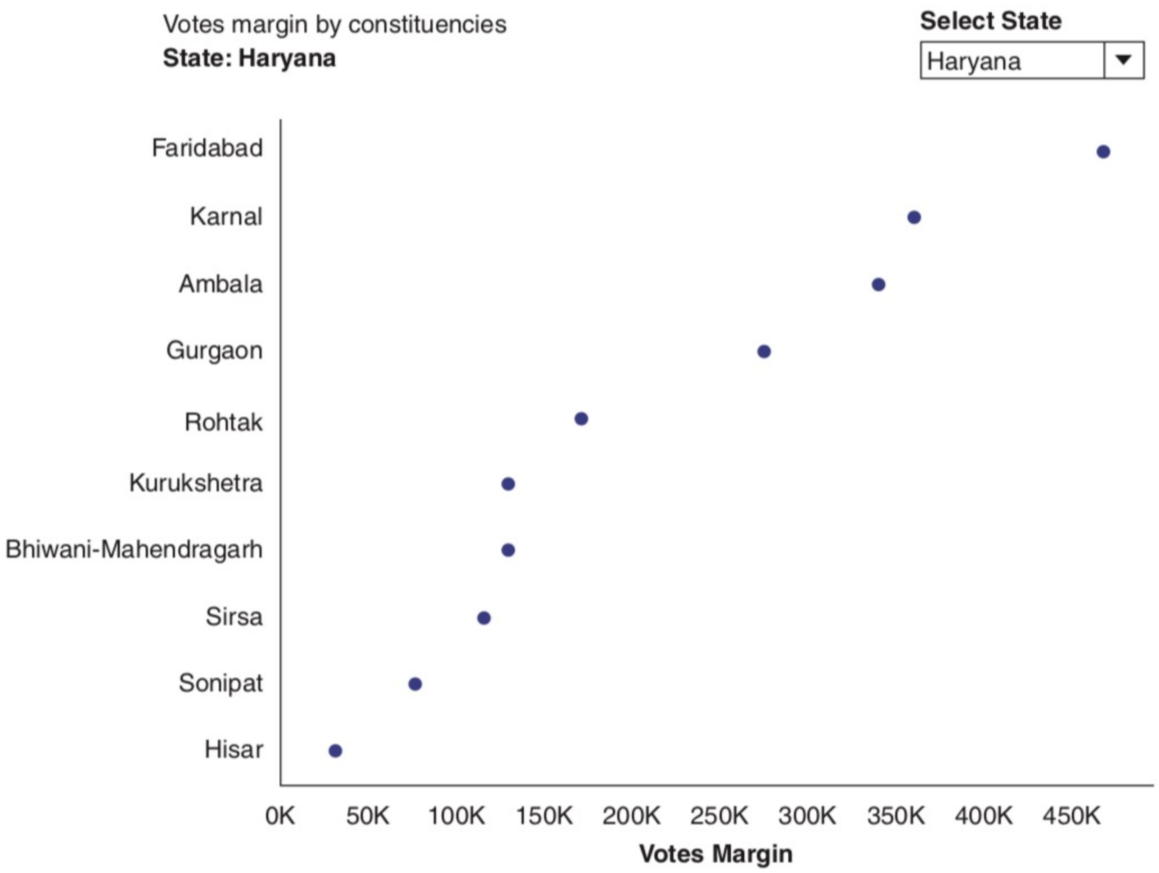
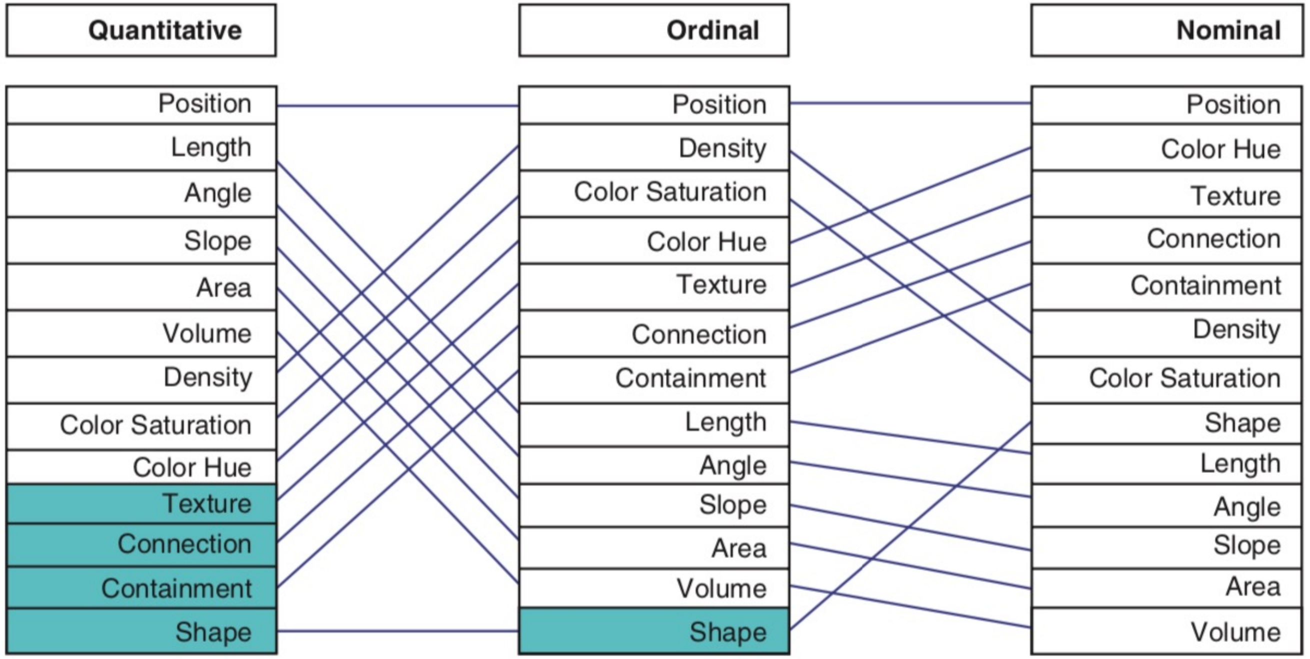


Figure 2.13 Winners' votes margin for all constituencies of "Delhi" for 2014 elections using position encoding.

- In the chart shown in Fig. 2.13, votes margin has been encoded as the position on a common scale.
- Using the chart, we can:
  1. **Discriminate between any two points.**
  2. **Since the data is sorted on the measure, ranking is trivial.**
  3. **Calculating ratio between any two points is easy too.**

Expressiveness and Effectiveness



Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Color Hue
Angle	Color Saturation	Texture
Slope	Color Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Color Saturation
Color Saturation	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

Figure 2.14 Graphical presentation of relational information. Ranking of perceptual tasks. The tasks shown in the grey boxes are not relevant to these types of data.

Source: Image is from Automating the Design of Graphical Presentations of Relational Information by Jock Mackinlay Stanford University.

- The design principles Expressiveness and Effectiveness for graphical presentation were defined by Jock Mackinlay in his thesis [Mackinlay, 1986].
- **Expressiveness:** This is about how well visualization shows all facts in a dataset.
  Effectiveness: Information is more readily perceived in a visualization compared to any other form.

Edward Tufte’s Design Principles

- Edward Tufte is seen as one of the most influential writers on the topic of visualizations.
- For charts and graphs, he provided graphical heuristic design principles on how to build efficient visualizations.
- Graphical design principles are a set of guiding “rules” about what works and what does not work in a design or composition.
- Tufte’s heuristic graphical design principles are used to enhance the readability of the graphs.

Principle 1: Maximizing Data-Ink Ratio

- Data-ink is the proportion of the ink dedicated to main elements of the chart, that is, quantifiable measures in the data.
- If data-ink is removed from the graphic, relevant information is lost
- On the other hand, non-data-ink refers to the ink which does not contribute to describing any information but is used in scaling and labelling of the graphic.
- Data-ink ratio [Tufte, 2001] is the non-erasable core of a graphic used for presenting the data.

$$\begin{aligned}\text{Data-ink ratio} &= \frac{\text{Data} - \text{ink}}{\text{Total ink used to print the graphic}}\\ &= \text{proprtion of a graphic's ink devoted to the non}\\ &\quad - \text{redundant display of data} - \text{information}\\ &= 1.0 - \text{proportion of graphic that can be erased}\end{aligned}$$

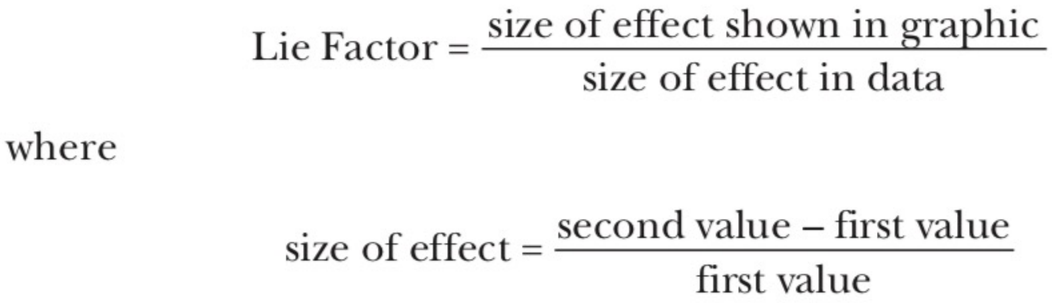
- A graphic is considered a good one when the data-ink ratio is close to 1.

Principle 2: Minimizing Chart Junk

- The unnecessary or confusing visual elements in a graphic are referred to as chart junk, which is not relevant in communicating any information.
- Edward R. Tufte [Tufte, 2001] has categorized chart junk into mainly three categories as follows:
  1. **Unintentional optical art:** The background of a data graphic should not be cluttered with unnecessary designs, colours, and optical art effects.
  2. **The grid:** The grid is used for initial plotting of data but not for printing, as grid lines can be counterproductive.
  3. **The duck:** Elements that focus on graphical style rather than the data are referred to as the duck by Edward R. Tufte. These are the elements that are put in the graphics for decorative purpose and focus on styling elements such as colours, shapes, back- grounds, and icons. These elements add no value in communicating the information our data contains.

Principle 3: Minimizing Lie Factor

- The “Lie Factor” [Tufte, 2001] is the ratio between the size of effect shown in a graphic and the size of effect shown in the data.



$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

where

$$\text{size of effect} = \frac{\text{second value} - \text{first value}}{\text{first value}}$$

- The value of Lie Factor should range from 0.95 to 1.05.
  - Any value below or above this indicates a substantial distortion in the graphic.
  - The graphic is called “overstating” if the value of Lie Factor is greater than 1 and “understating” if its value is less than 1 [Tufte, 2001].

Applying the Tufte's Principle in "Lok Sabha election analysis"

- Check the overall votes margin in 2014 for "Haryana"

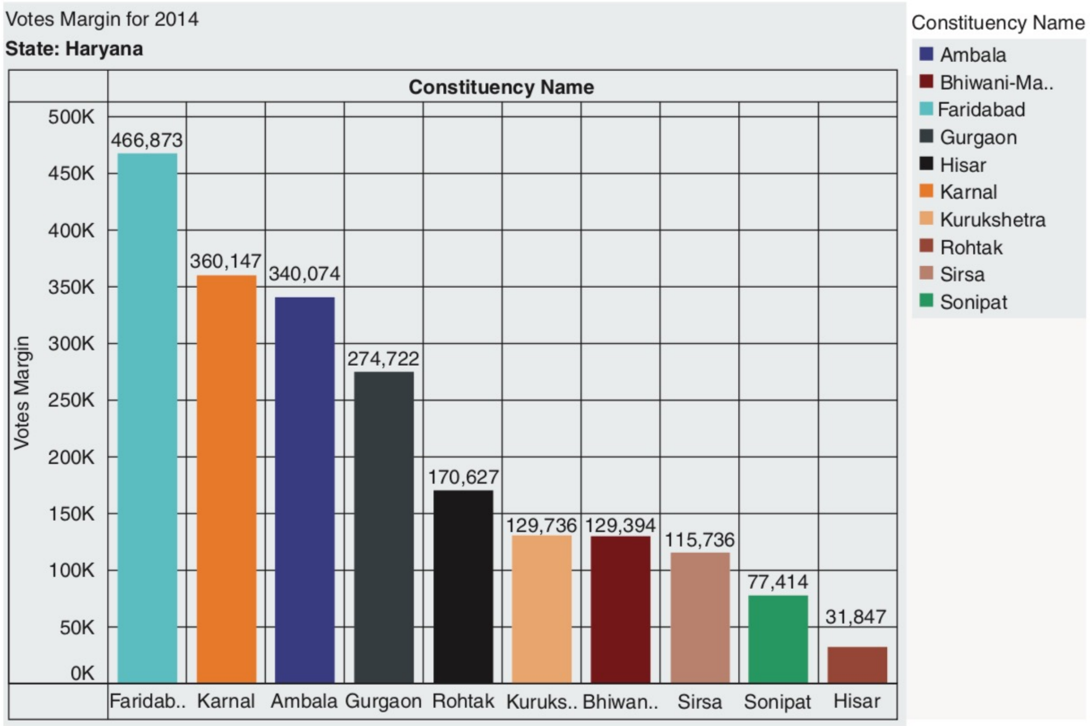


Figure 2.16 Winners' votes margin across constituencies of "Haryana" for 2014 elections.

1. We have a background colour in our charts. These do not add any information; hence we can clear the background.
2. The legend is not adding any value in our chart. Hence, we can remove the legend.
3. Since the bars are anyway labelled directly for constituency names, we do not need colour encoding. Hence, we can clear that too.
4. We can remove or soften gridlines to train the focus on the bars in the chart.
5. We have a descriptive title. We can update this to emphasize problem statement and remove axis labels.
6. We can emphasize on a particular value using annotation or colour coding.
7. Since the votes margin is specified on each bar, we can remove the *y*-axis scale.

Applying the Tufte's Principle in "Lok Sabha election analysis"

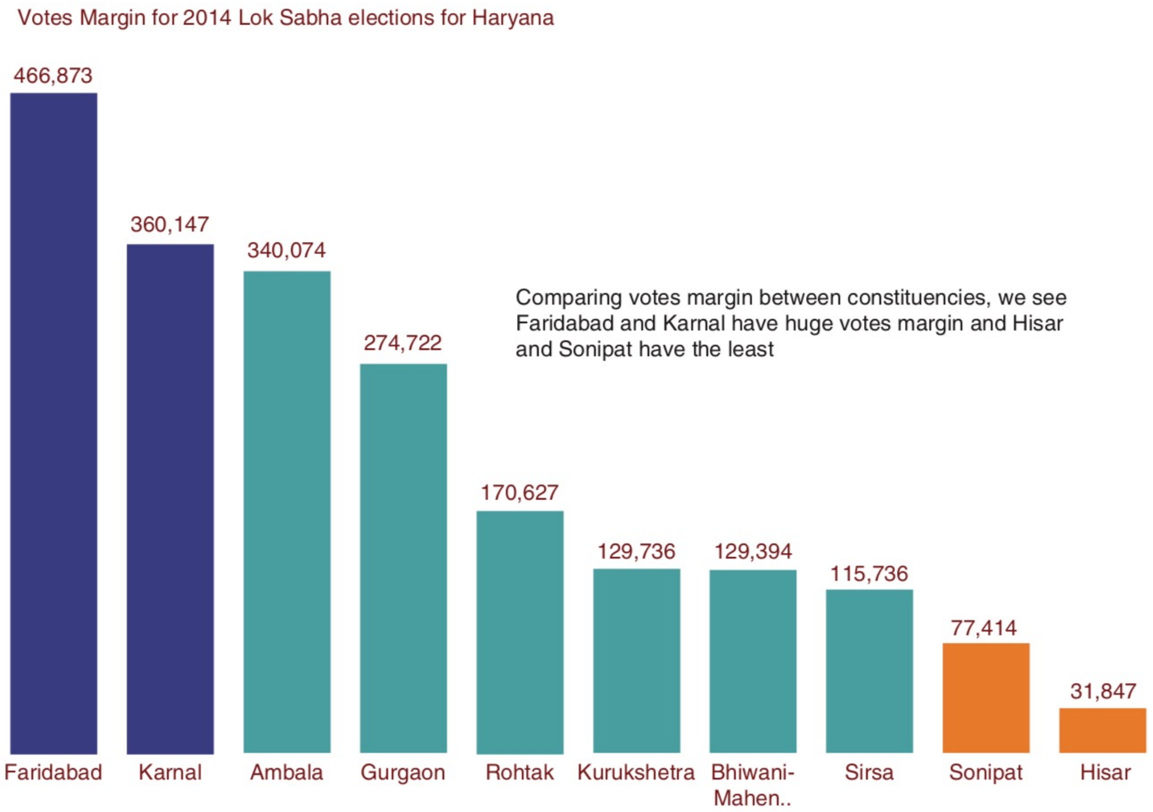


Figure 2.18 Winners' votes margin for constituencies of "Haryana" for 2014 elections.

1. The readability of our updated graphs is much better.
2. With these heuristic principles, we learnt that whenever we want to enhance the readability of visualizations, **consider what we can take away rather than what we can add.**

Can Chart Junk be Useful?

- As we understand Edward Tufte's graphical integrity principles, we need to avoid overindulgence on design and style that can suppress quantitative information.
- However there have been debates about the concept of chart junk and data-ink ratio.
- Nigel Homes a successful graphic designer challenged these concepts, his argument has been that by combining image with the graph, we can increase the memorability of a graph (Bateman, 2010)

Can Chart Junk be Useful?

- A study done by Scott Bateman [Bateman, 2010] measure the interpretation and recall of Holmes-style charts and plain visualizations with minimalist approach.
- The study concludes that visual embellishments may benefit the reader to a certain extent.
- So, does this mean that we load the chart design with embellishments?
- Certainly not, while we can see chart junk would be useful in bringing memorability to the charts, we need to consider that too much design variations would dis- tract the audience.

Can Chart Junk be Useful?

- A study done by Scott Bateman [Bateman, 2010] measure the interpretation and recall of Holmes-style charts and plain visualizations with minimalist approach.
- The study concludes that visual embellishments may benefit the reader to a certain extent.
- So, does this mean that we load the chart design with embellishments?
- Certainly not, while we can see chart junk would be useful in bringing memorability to the charts, we need to consider that too much design variations would distract the audience.

Edward Tufte’s design principles with slight variation

1. **Show comparisons** – Design of the chart should aid comparisons
2. **Show causality** – Showcase how variables impact or influence one another
3. **Use multivariate data** – Bring in varied sources of data to help your audience to interpret graph and understand the underlying narrative
4. **Completely integrate modes** – Use multiple modes of information such as texts and calculations to explain the details to your audience
5. **Establish credibility** – Provide details on data source, scales, measurements to establish credibility on your narrative with your audience
6. **Focus on content** – Show data with the maximum data-ink ratio. Erase redundant/non-data-ink, within reason

References

- [Cooney, 2018] – M. Cooney (2018), “Cisco Predicts Nearly 5 Zettabytes of IP Traffic per Year by 2022”, *NetworkWorld*, available at https://www.networkworld.com/article/3323063/cisco-predicts- nearly-5-zettabytes-of-ip-traffic-per-year-by-2022.html, last accessed May 27, 2021.
- [Anon, 2020] - Anon (2020), Domo Releases Eighth Annual “*Data Never Sleeps*”, available at https://www.domo.com/news/press/domo-releases-eighth-annual-data-never-sleeps-infographic, last accessed May 27, 2021.
- [Cleveland,1985]–W.S.Cleveland(1985),W.ClevelandandR.Mc-Gill (1985), “*Graphical Perception and Graphical Methods for Analysing Scientific Data*”, *Science*.
- [Cleveland, 1994] – W. S. Cleveland (1994), Cleveland, and W.S. (1994). *The Elements of Graphing Data*. 2ed. Hobart Press, Summit, NJ, USA. 323 pp.
- [Mackinlay, 1986] - J. Mackinlay (1986), “*Automating the Design of Graphical Presentations of Relational Information*”. ACM Trans Graph 5. 110–141. 10.1145/22949.22950.
- [Bateman, 2010] - S. Bateman (2010), “Useful Junk? *The effects of visual embellishment on comprehension and memorability of charts*”, Source: DBLP, DOI:10.1145/1753326.1753716.
- [Tufte, 2001] – E. Tufte (2001), Design Principles, *the Work of Ed- ward Tufte and Graphic Press*, available at https://www.edwardtufte. com/tufte/, last accessed June 19, 2020.
- [Bateman, 2010] – S. Bateman (2010), “Useful Junk? *The effects of visual embellishment on comprehension and memorability of charts*”, Source: DBLP, DOI:10.1145/1753326.1753716.

Thank You!