INDIAN INSTITUTE OF TECHNOLOGY MADRAS

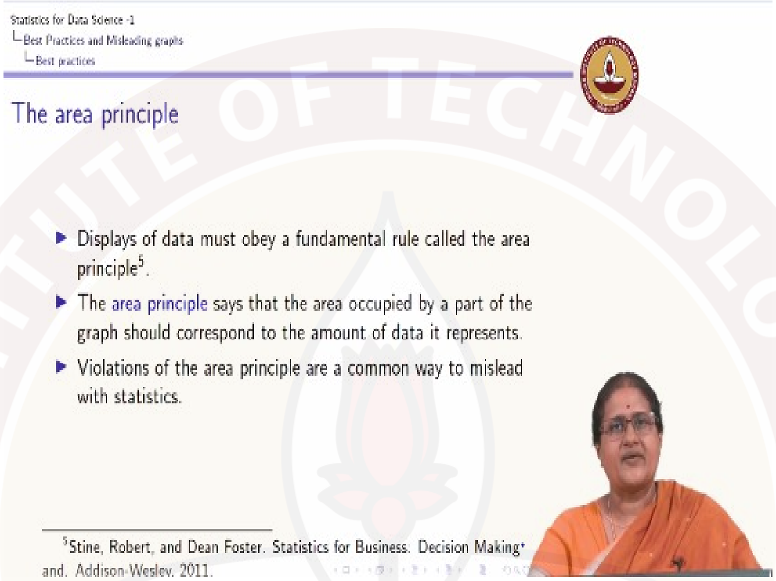
सिद्धिर्भवति कर्मजा

IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.4
Describing Categorical Data - Best Practices While Graphing Data - 2

(Refer Slide Time: 00:14)



Statistics for Data Science -1
└Best Practices and Misleading graphs
  └Best practices

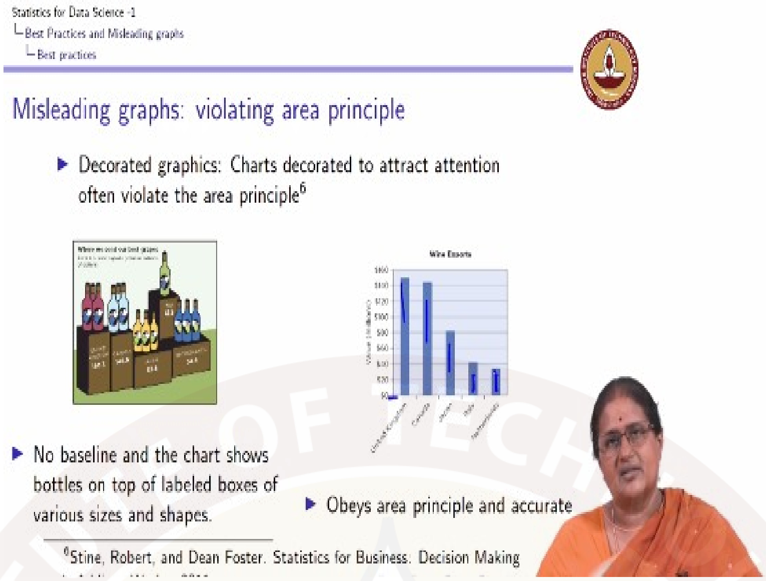
The area principle

- Displays of data must obey a fundamental rule called the area principle[5].
- The area principle says that the area occupied by a part of the graph should correspond to the amount of data it represents.
- Violations of the area principle are a common way to mislead with statistics.

[5] Stine, Robert, and Dean Foster. Statistics for Business. Decision Making and. Addison-Wesley. 2011.

Now whenever we come to display a data there is a fundamental rule called the area principle. Now, what is this area principle actually tell us. When we look at the area principle the area principle states that the area occupied by a part of the graph should correspond to the amount of data it represents. What do we mean by this?

(Refer Slide Time: 00:42)



Statistics for Data Science -1
└Best Practices and Misleading graphs
└Best practices

Misleading graphs: violating area principle

- Decorated graphics: Charts decorated to attract attention often violate the area principle[6]

- No baseline and the chart shows bottles on top of labeled boxes of various sizes and shapes.
- Obeys area principle and accurate

[6] Stine, Robert, and Dean Foster. Statistics for Business. Decision Making

You can look at the first thing is when I have charts which are decorated to attract attention. For example let us look at this particular chart now what this chart gives us is the total wine exports in the United States value in millions of dollars. Now this is more like a infographic, but what does this tell? It tells us United Kingdom is about 150.3, Canada is 146.8, Japan is 82.8, Italy is 42.5 and Netherlands is 34.4.

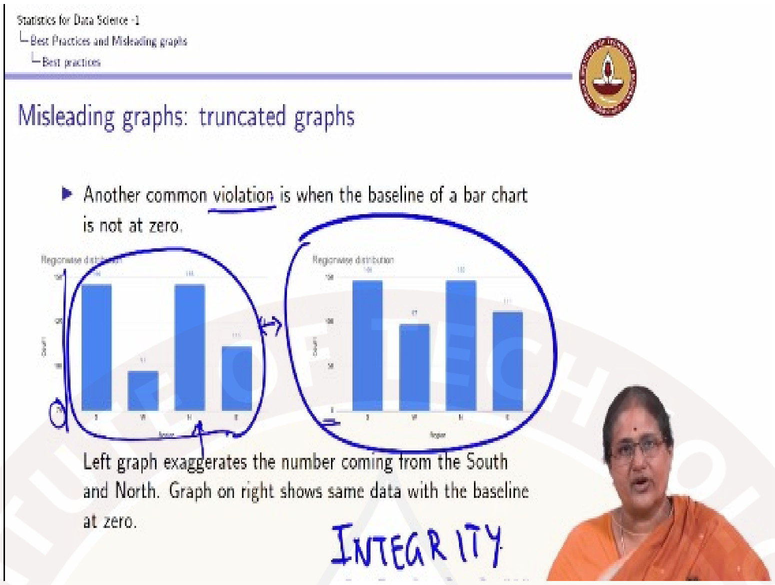
You can immediately see that there is no baseline, it show bottles on top of label boxes. Now even the boxes are not of any uniform shape they are of various shapes and sizes. So, when you look at this chart it does not convey anything which is actually I can make a meaning out of. Whatever I need to convey is just a total US wine exports to each one of these countries and I can construct this using a bar chart where my categories are again United Kingdom, Canada, Japan, Italy and Netherlands.

I have labeled each one of my categories. I can see that United Kingdom has a value in 150.3 million dollars, Canada is about 146.8 million dollars, 82.8 to Japan, 42.5 to Italy and 34.4 to Netherlands. Now the chart on the right hand side obeys what we call the area principle. It is accurate, it has a baseline this is the baseline of this chart. So, you can see that this is the baseline of this chart.

It has an everything is it is actually consistent the width of the bars for each countries is equal and I can have a vertical scale and on the vertical scale I have the value which is given and hence

I obeys what I refer to as the area principle where the area occupied by the graph and what is the area occupied by the graph it is this area, it is this area, it is this area, it is this area and this area it is proportional to the data that is being presented.

(Refer Slide Time: 03:21)



Statistics for Data Science -1
└Best Practices and Misleading graphs
└Best practices

Misleading graphs: truncated graphs

- Another common violation is when the baseline of a bar chart is not at zero.

Left graph exaggerates the number coming from the South and North. Graph on right shows same data with the baseline at zero.

INTEGRITY

The next way people mislead with graph is through use of what we call as truncated graph. Now what is a truncated graph? Now let me show one thing is where the baseline of a bar chart is not at 0. What do we mean by baseline of a bar chart is not at 0? Now let me show you this graph now again this is a regional wise distribution. I have four regions South, West, North and East.

Now when I portrayed this data and this graph to you, you can immediately see there are 500 people and I have just the data is which region do the 500 people come from, that is the data. Now when I portray or I show this graph the immediate thing or the immediate response or a person who looks at this data without looking at the access is to imagine that these people from South and North are much higher than the people from West or East.

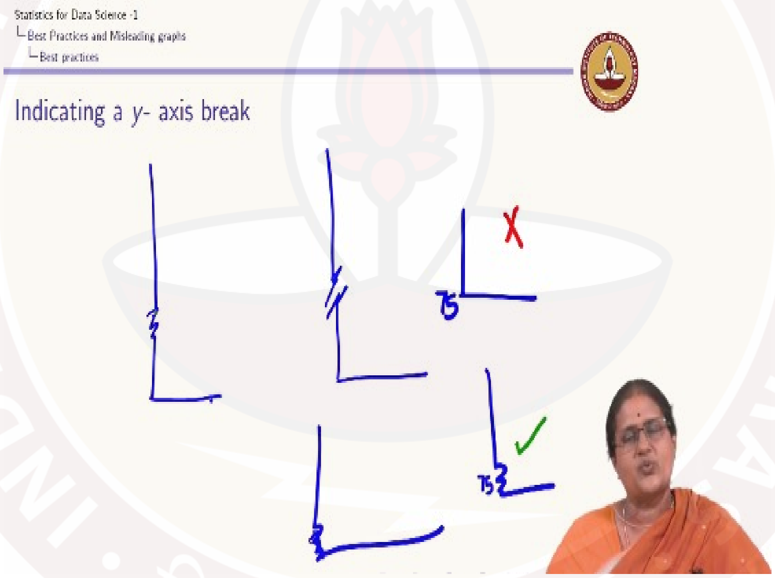
That is what this message this graph conveys, but when we look more carefully at the data you see that this is the baseline and it starts at 0 whereas here it starts at 75. Now these two graphs are actually the same data. What do I mean by same data? 500 students each one of them tell which state, region they come from whether it is a South, whether it is a West, whether it is a North or the East.

When you look at this graph, you feel that the South and North are the ones from where majority of the student come. Even though majority come it seems that the North and East are negligible to the distribution from North and South, whereas this graph conveys a different story. So, the left graph or this graph exaggerate the number whereas this graph shows the same baseline at 0.

Now, this is what is recommended because this is where data integrity is maintained and the actual story is said. Now where does this matter, now when you are showing a growth and everything and you truncate the graph or you miss the present data or you mislead, people will actually attribute a wrong story to this even though both of them represents the same data. Visually you see that this tells a different story than what do the one on the right says.

So, the second thing is whenever you truncate graph there is a loss in information and it is a violation this has to be avoided.

(Refer Slide Time: 06:42)



Statistics for Data Science -1
└Best Practices and Misleading graphs
  └Best practices

Indicating a y- axis break

Now some textbooks and some people say that whenever you have a truncated graph or something you introduce a y axis break. So, if you are starting and you are breaking the axis you can introduce it either by doing this where I am telling that I am introducing a break in my y axis or you can show that this is I am not starting from 0 there is a break and I am starting from a higher value.

For example, if I wanted to start from 75 I should have indicated that I am starting from 75 instead of shifting the graph to 75. So, this is incorrect whereas this is this is the right way to go about it. So, whenever you are altering or manipulating with the y axis, indicate to the reader that you have manipulated or you have introduced a y axis break.