INDIAN INSTITUTE OF TECHNOLOGY MADRAS

सिद्धिर्भवति कर्मजा

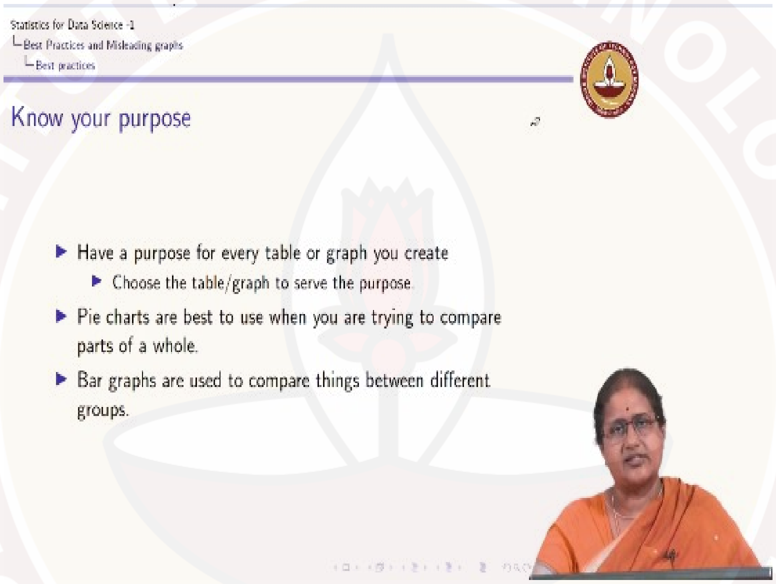
IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.3
Describing Categorical Data - Best Practices While Graphing Data - 1

So, what we are going to do now is to understand certain best practices about drawing these pie charts and bar charts what we have learned so far.

(Refer Slide Time: 00:24)



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

Know your purpose

- Have a purpose for every table or graph you create
  - Choose the table/graph to serve the purpose
- Pie charts are best to use when you are trying to compare parts of a whole.
- Bar graphs are used to compare things between different groups.

Now even before we actually go and understand about what is required from a bar chart and pie chart or a frequency table, the first thing we need to understand is what is the purpose? Now when I said purpose it does not mean or it does not necessarily mean that every dataset should convey a message. What is the purpose? We asked our question what is the purpose? For example if I just have a data say I have a data which is A, B, C, A, B, D, E, A or D, A, B, B this is just a categorical data.

Now if I have this data, now the questions I need to ask is what is the purpose of this data? Suppose this data is just a set of states and nothing more than that suppose this is Andhra Pradesh, this is Bihar, this is Chhattisgarh and all of that and I am looking at I am collecting data or I am asking every person who is entering a particular room which state they belong to and this is the data I have.

Now once I collect this data I need to understand what are the questions I want to ask from the data? The first thing is if I want to just tabulate this data. So, I look at it, I prepare a table like the way we discussed in the last time I say people from state from A, people from state B, people from state D, people from state D this is the category I am looking at and then after I have the count which I called it is also the frequency.

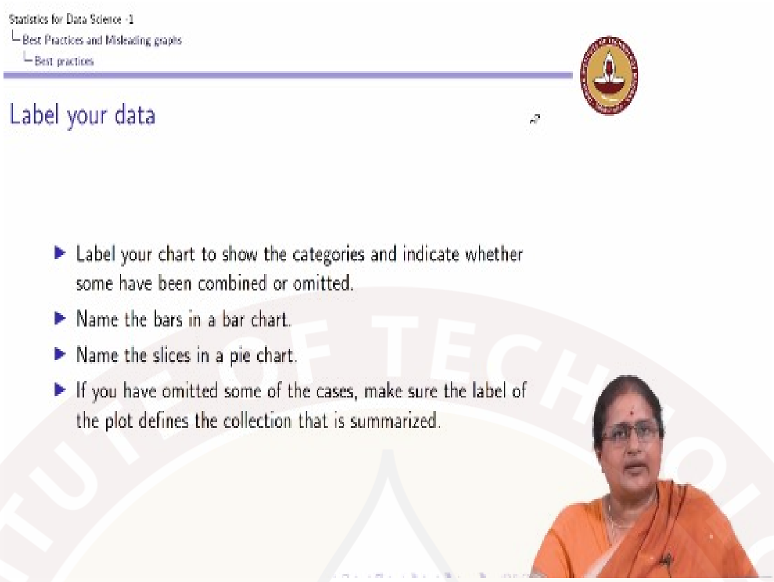
Once I have this count this could be say 1, 2, 3, I have 3 I have 1, 2, 3, 4, 4. C is just 1 and D is 2. So this is the way I construct what is called a frequency table. So, now the question is what is the purpose from this if my purpose is just to count and represent it as a table I go in for a frequency table. However, if my purpose is to come up with a tabulation I want to compare how each state does with the other.

Then I might want to go for a bar chart because bar chart helps me in comparison. However, if my purpose is to know what is the share of each state then I will go in for relative frequencies which is depicted by a pie chart. So, what we are going to do today is to first understand when to use a pie chart. You use a pie chart when we are trying to compare parts of a whole.

So, if I have class of say 100 students and I want to know what is the regional distribution, then after a pie chart is the right chart for me to use to compare this. However, if I want to compare things I want to know what is the exact count of how many are from each region then a bar chart is more appropriate. So, the first thing when would you use a table. We would use a table when we are interested.

Suppose I have a lot of categories then what would happen is both the bar chart and pie chart would appear to be entirely clattered. If I want to represent the entire data then perhaps a table is more appropriate. So, even before we go to summarize data the first question we need to ask is what is the purpose of my summary, analysis, what is it I want to convey. Choose the table or graph to serve this purpose.

(Refer Slide Time: 04:20)



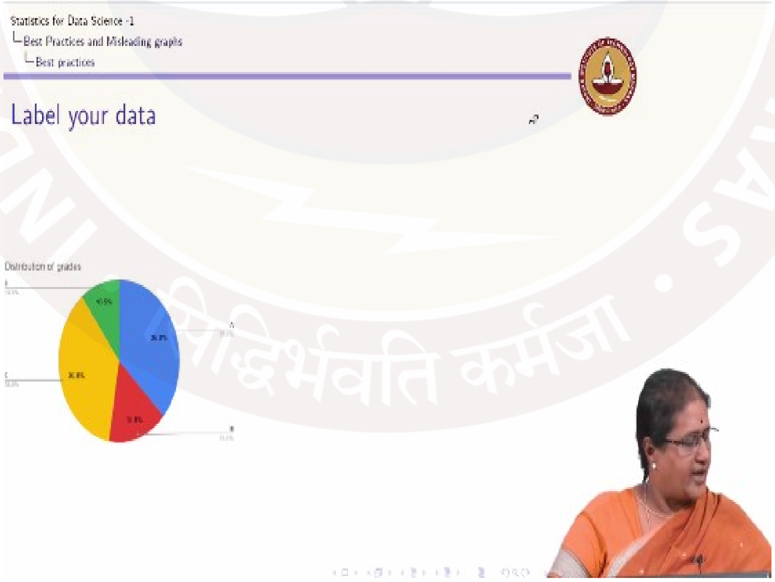
Statistics for Data Science -1
└Best Practices and Misleading graphs
　└Best practices

Label your data

- Label your chart to show the categories and indicate whether some have been combined or omitted.
- Name the bars in a bar chart.
- Name the slices in a pie chart.
- If you have omitted some of the cases, make sure the label of the plot defines the collection that is summarized.

The next very important thing is what we saw in the earlier class was how to create both the bar chart and the pie chart. Now we are going to spend some time to understand how to label the bar charts or how to label the charts.

(Refer Slide Time: 04:41)

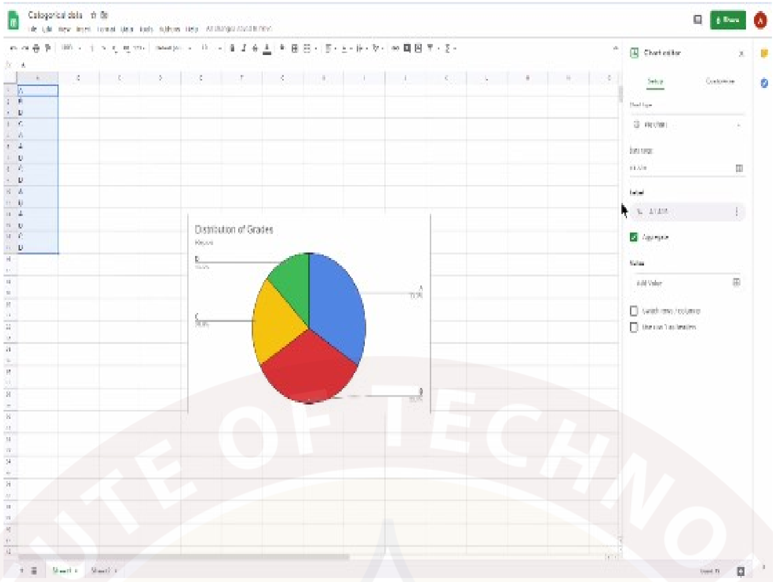


Statistics for Data Science -1
└Best Practices and Misleading graphs
　└Best practices

Label your data

So, even before we label the chart.

(Refer Slide Time: 04:43)



So, now let me just go back to our Google sheet and type in a particular data. Suppose, this is the data I have just typed in I just want to create a chart of this data, so I go to insert, I go to chart and you can see that I have a pie chart which is inserted here. Now within the pie chart you can see that I have a tab which says it is customized and that is on your right hand corner. Now we can see that, here this pie chart I have what are the pie slices.

I know I can put up a title this title is something currently the title is count I can just put a title which says distribution of some grades. If A, B, C, D represents grades I can just say distribution of grades. Now typically if you want a subtitle you can add a subtitle. For now I am just adding a subtitle which is region just for the sake of it and you can see the distribution of grades region appears here.

Now it always helps us to then look at what are the legends you can want you can put the legend wherever you want and I am sure that we will have a separate tutorial to actually tell how different legends would look later, but at this point of time what I want you to tell is you can see what is the color whether it is auto and you can keep adding the colors and the text colors or you can just put it auto the title font size.

You can look at what are the chart this one and suppose I put 25 you can see that you are actually moving a particular slice out of the particular thing. One point I want to make here what I want to clarify is when you are clicking two colors is not advisable because when you click two colors it is really you do not have any change between these two colors. So it is always advisable to retain one particular color.

Now suppose I do the same thing for each of these pie slices I go and I do each of them. This even though it is possible you see that this does not make any sense. The reason why a chart like this does not make any senses it is not actually the region the purpose of a pie chart is to help you understand what is the share of a pie. Now when just for decorative purposes if people start demonstrating it this way you can see that this does not actually convey what you want to say.

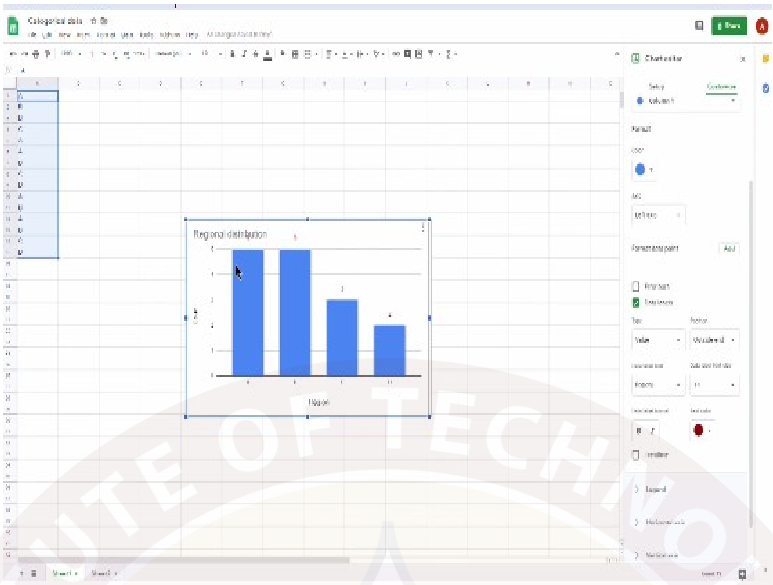
So I would strongly advise against a visualization where you are actually moving every pie slice rather I would prefer that unless essential do not have to change the distance from the center and have everything maintained so far we have a nice pie chart. Again this would tell us how to label it whether you want it in the label what is the font size and whether you want it as a in italics or whether you want it bold you can see as I keep changing here.

So, all of that you can do it later. Now another thing is if you want borders you can add a border or adding a border would always help us. You can always decide whether you want a border or not. There is something else in the layout which is called a 3D. Again, I do not think that 3D is giving us any extra value, but you can always have a pie chart as a three dimensional chart also.

You use it whenever you want it, but in my opinion a simple pie chart conveys more than what is essential. So, this is about how you are going to come up with chart title, chart subtitle and what are the default size for each pie slice you can see what is the distance from the center. You can customize it and then after you can add values whenever you want, you can switch rows and columns here but here since I have only one column I do not have to do that.

Now the same thing, the same data I can choose here for a same data I can choose the chart which I call a bar chart or a column chart here.

(Refer Slide Time: 10:12)

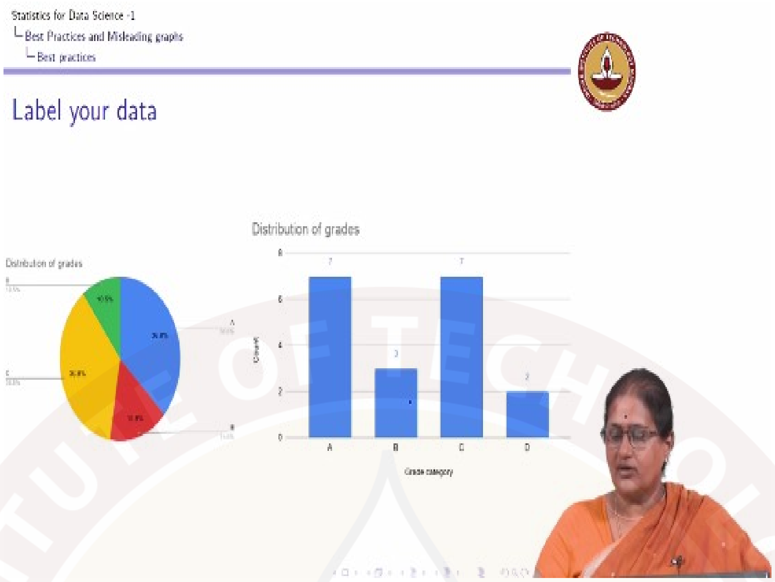


Now again in a column chart I can customize the column chart. First thing I add the subtitle and title, I can add the chart title again I am going to add the chart title as regional distribution assuming these are A, B, C, Ds are regions. I can add a subtitle if I want to. Here I need to add a horizontal axis title. In the horizontal axis title I can just add region because my A, B, C, D are actually representing regions.

Again within a region I can tell what is the font I needed if I need a very big font size and if I need it in bold I can do all those things here. I also have a vertical axis, in the vertical axis I just add a count. The vertical axis this count is given. So, this count is given 1, 2, 3, 4, 5 so another thing which I would like to add is I can add data labels where data labels it tells me actually what is the number of counts or observations for each category.

Now this auto labels can be in any size, too big or it can be 5. The text color I can choose the text color I need for my data labels and I also can tell what is the position. I always prefer it at the outside end or I can have it at the inside base wherever you want. This I am not going to have there is no particular actual to be done it or this is what you can have, you can see that. You have the region with the distribution which is at the outer end.

(Refer Slide Time: 11:55)



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

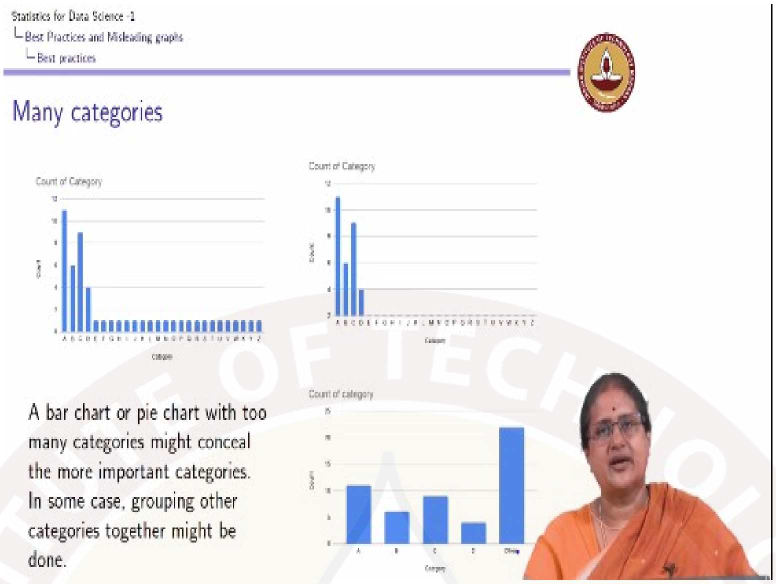
Label your data

So, what the first thing which we need to understand is to be it a pie chart or a bar chart. First thing is to label or annotate your data because only when we label or annotate the data there is a better visualization or it communicates the idea better. So, here again if these were A, B, C, D I have label my chart distribution of grades and this gives me the distribution of grades which is 36.8 %% of A 15.8 % of B, 36.8% of C and 10.5% of D.

Similarly, this is giving me the count. I have 7 people who have got A, 7 people who have got C whereas 3 people who have got B and 2 who have got D. So, you can see that this is telling the share of a particular grade whereas this is giving a count. So, whenever you want to represent data in either a pie chart or a bar chart the first thing you do is label or annotate your chart.

Now what you see the horizontal lines here these are called gridlines. Now you can also choose whether to have the grid lines or not that you can go and you can choose again in your data whether you would want to have gridlines or not and that can be done here whether you want the grid lines or not you can actually choose it here. So, when I have the option of no gridlines I get a chart without gridlines. Now suppose I am giving you another data set.

(Refer Slide Time: 13:53)



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

Many categories

A bar chart or pie chart with too many categories might conceal the more important categories. In some case, grouping other categories together might be done.

Now if you look at this data set it tells me that I have about 11 of category A, 6 of category B, about 9 of category C and about 4 of category D and I have many, many, many, many more categories each one of them they have only one of each such category. Now where could such data come up from? Now suppose for example I am asking some hundred people who is your favorite cricketer?

There would be an overwhelming response that Sachin Tendulkar is their favorite cricketer, some of them might choose Virat Kohli others might choose say Zaheer Khan some might choose MS Dhoni, but then afterwards you have all these little, little guys only two of them might choose say Rahul Dravid, two might choose KL Rahul, two might choose Ashwin and all of this.

So, this is in a sense that the major choice could be the first 4 cricketers and others it is not that my entire 100 people are going to only choose among these 4 cricketers, but I have a distribution. So, in this case what you see is when I immediately look at a graph of this kind I find too clattered because there are too many categories. Now this could also be a case where I am looking at distribution of industries in particular states.

There could be only 4 states which sees maximum number of industries and I have a splatter of industries in other states. So, now we have too many categories and the bar chart actually looks very clattered and it is not conveying what I am looking for it to convey. What do we do in these cases? So, a bar chart or pie chart with too many categories might actually conceal the more important categories.

So, one way to do it is do not ignore the category. So, the chart now currently what chart I have not plotted those categories I have only given a command that plot the bars only if it has more than 2 count. So, you can see that all my categories have been which had lesser than 2 counts are not figured in the bar chart at all. What I would suggest is go to a category here where the others are clubbed into a major category.

Now if you do this it conveys two important things. One is you are not excluding any data and the second important thing which it conveys is that even though this bar chart says 11 of the total come from category A and says that 9 from category B what this chart conveys is of the total number you have a significant number that comes from the smaller categories and that you can see is about 22.

So, this gives so eliminating or not giving this information actually does not give you the entire story. So, this is one way where you can club in all the smaller categories or categories which have very little count to call it as an other category and this when you portray many categories as an other category it gives you the entire story.