Describing a categorical variable

Describing a categorical variable

Suppose you ask 50 people to pick which color they like best among green, blue, and red.

Describing a categorical variable

Suppose you ask 50 people to pick which color they like best among green, blue, and red.

Then your variable is categorical, and your observations might start out like this:

blue red red green red green red red red blue . . .

Describing a categorical variable

Suppose you ask 50 people to pick which color they like best among green, blue, and red.

Then your variable is categorical, and your observations might start out like this:

blue red red green red green red red red blue . . .

A natural first step towards a summary of such a list is to count the number of each kind:

Describing a categorical variable

Suppose you ask 50 people to pick which color they like best among green, blue, and red.

Then your variable is categorical, and your observations might start out like this:

blue red red green red green red red red blue . . .

A natural first step towards a summary of such a list is to count the number of each kind:

color	**number of people**
red	30
blue	10
green	10

Describing a categorical variable

Suppose you ask 50 people to pick which color they like best among green, blue, and red.

Then your variable is categorical, and your observations might start out like this:

blue red red green red green red red red blue . . .

A natural first step towards a summary of such a list is to count the number of each kind:

color	**number of people**
red	30
blue	10
green	10

This simple summary tells you a lot about the people's preferences.

Describing a categorical variable

Suppose you ask 50 people to pick which color they like best among green, blue, and red.

Then your variable is categorical, and your observations might start out like this:

blue red red green red green red red red blue . . .

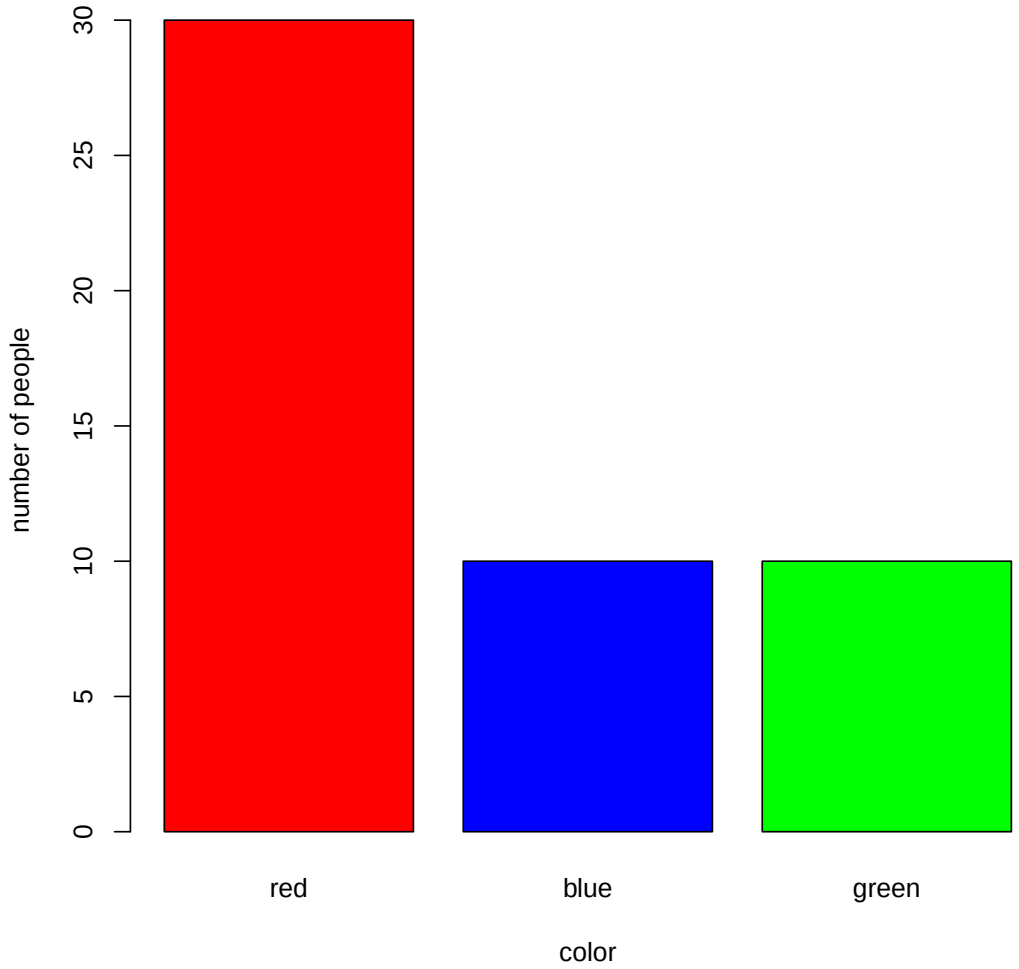
A natural first step towards a summary of such a list is to count the number of each kind:

color	**number of people**
red	30
blue	10
green	10

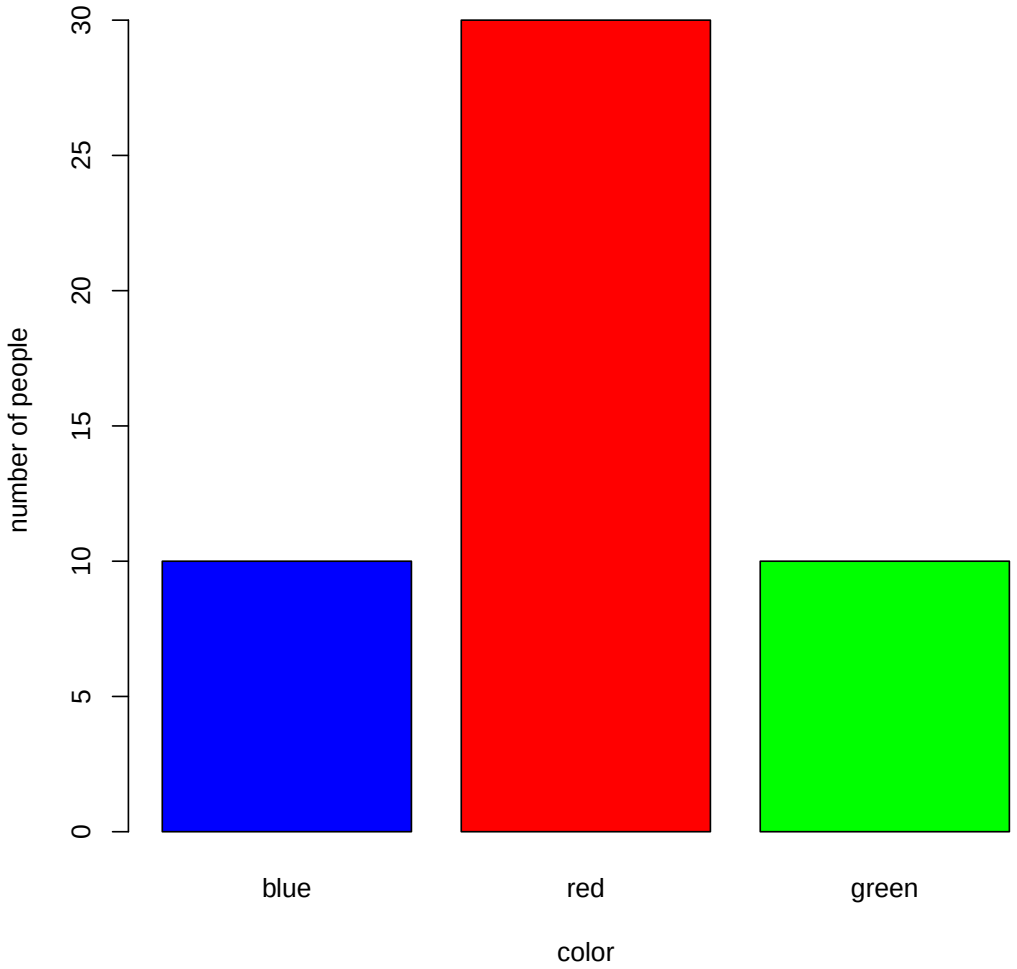
This simple summary tells you a lot about the people's preferences.

A picture is more vivid. Here is a common way of graphically describing a categorical variable.

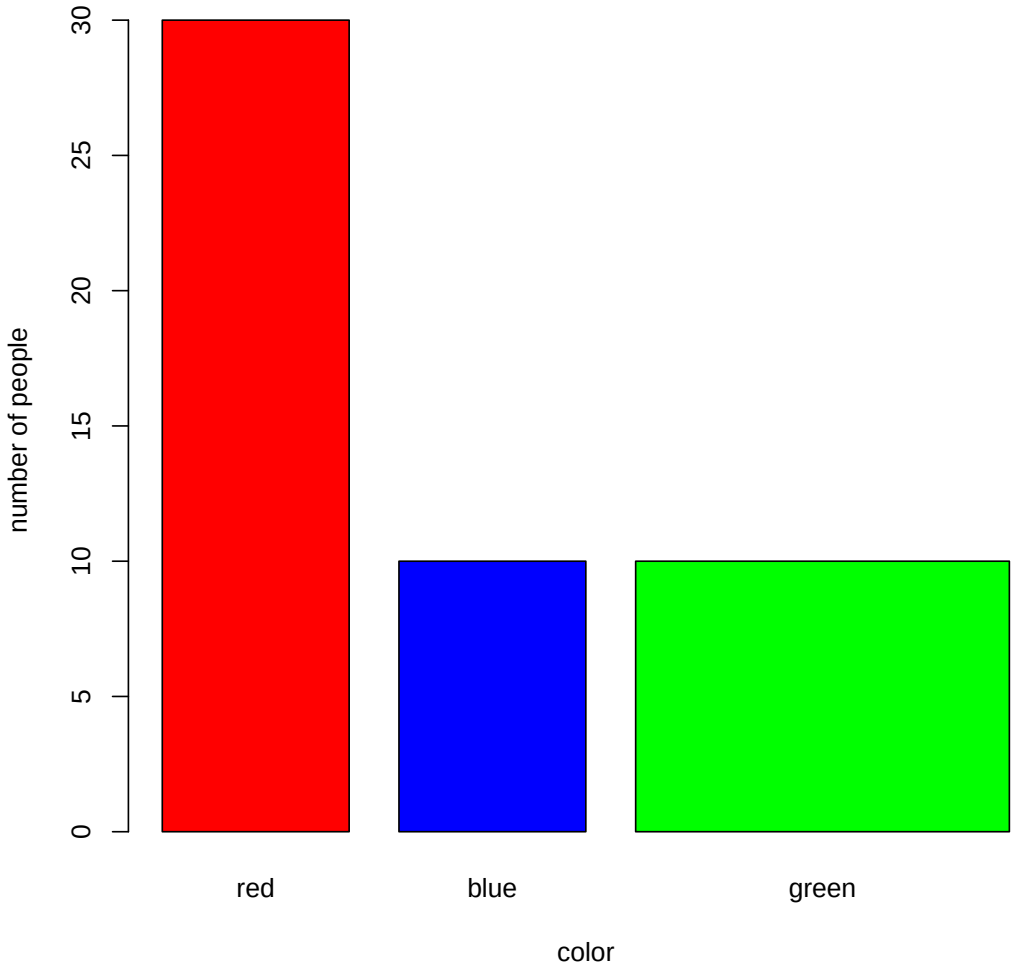
Bar graph of the categorical variable "favorite color"



Bar graph: favorite color, different order



Bad bar widths!!!



Areas matter

The **area** of a bar is visually important.

Areas matter

The **area** of a bar is visually important.

Keeping the width of the bars equal ensures that not only the height but also the **area of each bar is proportional to the number of people in that category**.

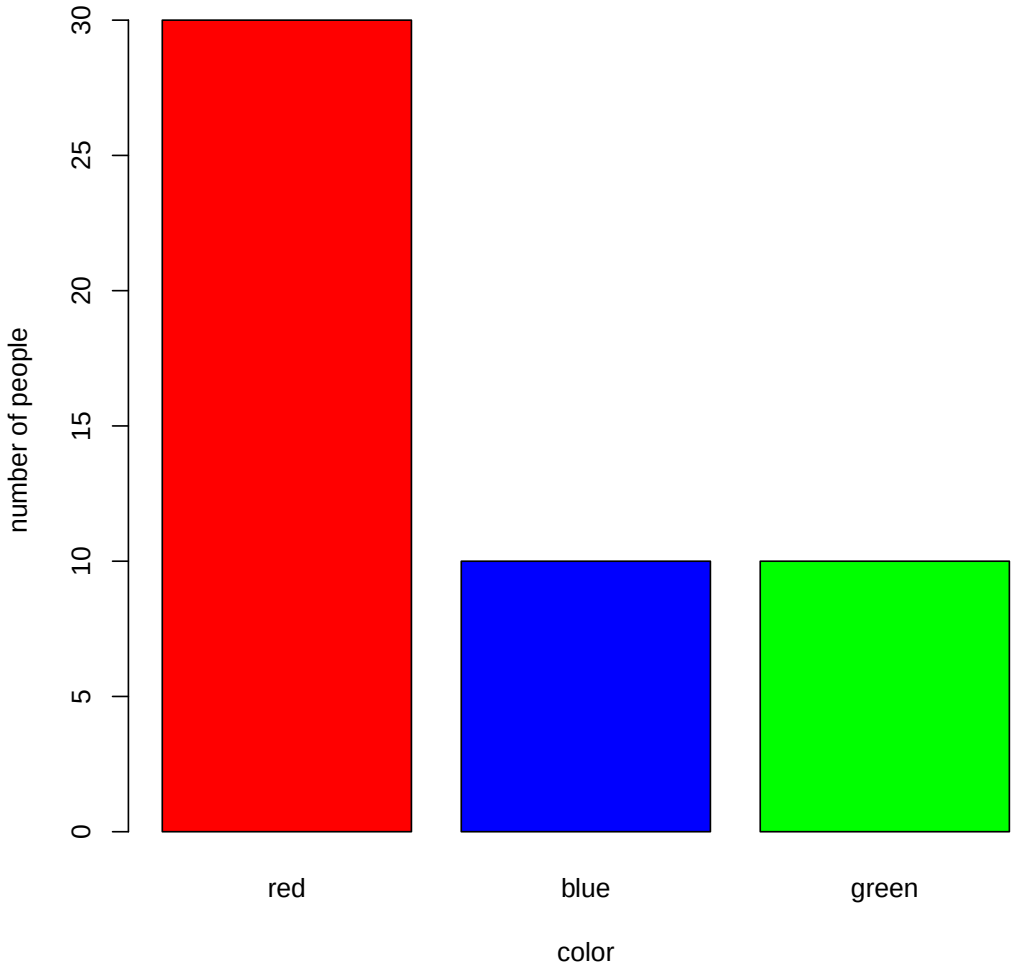
Areas matter

The **area** of a bar is visually important.

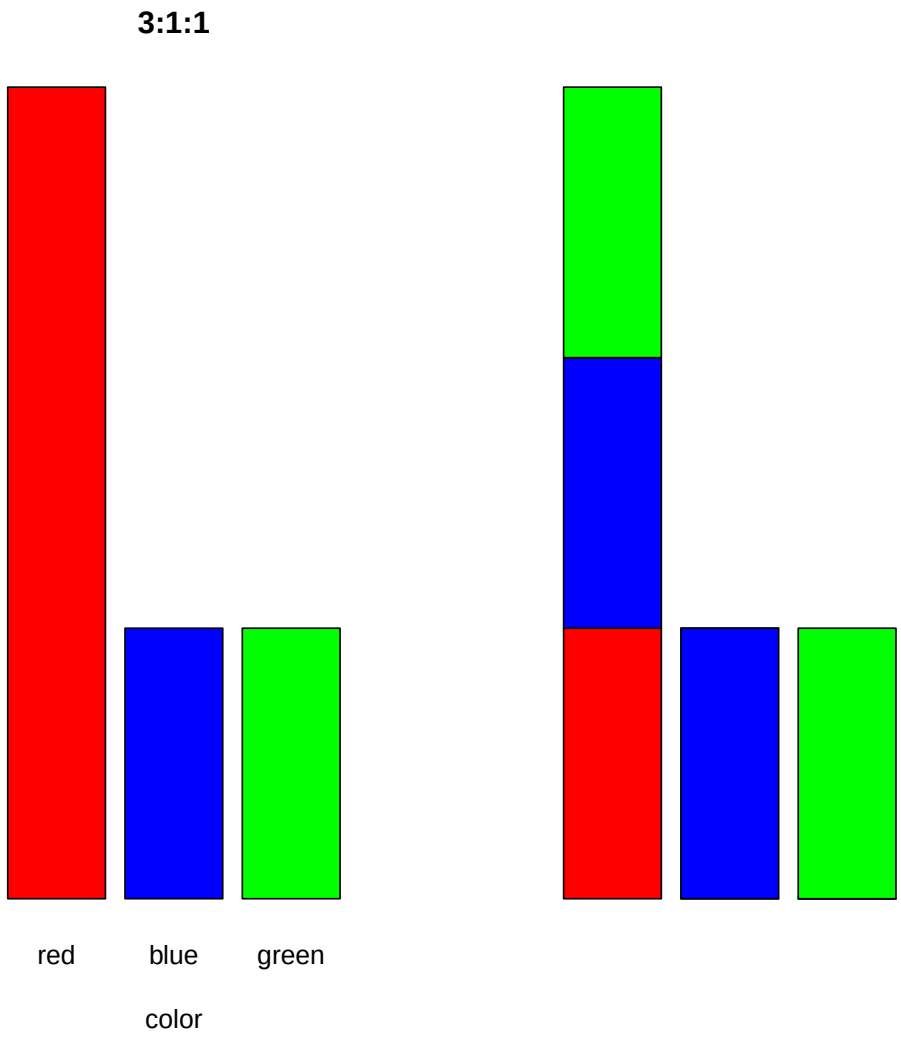
Keeping the width of the bars equal ensures that not only the height but also the **area of each bar is proportional to the number of people in that category**.

This observation will be crucial when we start creating graphical summaries of quantitative variables.

A good bargaph



No vertical axis needed; can still see relative proportions



Bar graphs and percents

Even without the vertical scale, you can compare areas and clearly see that the colors red, blue, and green appear in the ratio 3:1:1.

Bar graphs and percents

Even without the vertical scale, you can compare areas and clearly see that the colors red, blue, and green appear in the ratio 3:1:1.
This means that you can think of the data being split into $3 + 1 + 1 = 5$ equal pieces, 3 of which are red, 1 is blue, and 1 is green. In other words:

Bar graphs and percents

Even without the vertical scale, you can compare areas and clearly see that the colors red, blue, and green appear in the ratio 3:1:1.
This means that you can think of the data being split into $3 + 1 + 1 = 5$ equal pieces, 3 of which are red, 1 is blue, and 1 is green. In other words:

color	**proportion of people**	**percent of people**
red	3/5	60%
blue	1/5	20%
green	1/5	20%

Bar graphs and percents

Even without the vertical scale, you can compare areas and clearly see that the colors red, blue, and green appear in the ratio 3:1:1.
This means that you can think of the data being split into $3 + 1 + 1 = 5$ equal pieces, 3 of which are red, 1 is blue, and 1 is green. In other words:

color	**proportion of people**	**percent of people**
red	3/5	60%
blue	1/5	20%
green	1/5	20%

So, without the vertical scale, you can recover all the information that was in our earlier summary, except the total number of people (which was 50).

Bar graphs and percents

Even without the vertical scale, you can compare areas and clearly see that the colors red, blue, and green appear in the ratio 3:1:1.
This means that you can think of the data being split into $3 + 1 + 1 = 5$ equal pieces, 3 of which are red, 1 is blue, and 1 is green. In other words:

color	**proportion of people**	**percent of people**
red	3/5	60%
blue	1/5	20%
green	1/5	20%

So, without the vertical scale, you can recover all the information that was in our earlier summary, except the total number of people (which was 50).

Apart from the numbers on the vertical axis, the appearance of the bar graph depends only on percents. It would have looked the same if the data consisted of 300 red, 100 blue, and 100 green;

Bar graphs and percents

Even without the vertical scale, you can compare areas and clearly see that the colors red, blue, and green appear in the ratio 3:1:1.
This means that you can think of the data being split into $3 + 1 + 1 = 5$ equal pieces, 3 of which are red, 1 is blue, and 1 is green. In other words:

color	**proportion of people**	**percent of people**
red	3/5	60%
blue	1/5	20%
green	1/5	20%

So, without the vertical scale, you can recover all the information that was in our earlier summary, except the total number of people (which was 50).

Apart from the numbers on the vertical axis, the appearance of the bar graph depends only on percents. It would have looked the same if the data consisted of 300 red, 100 blue, and 100 green; or 4800 red, 1600 blue, and 1600 green; and so on, as long as the percents were 60, 20, and 20.