# Associations in Data

# Accelerated 7
# Unit 6

## 8 Synthesis

Choose one representation that is used to organize data. Describe its advantages.

- List
- Table
- Scatter plot

## Summary

Data that includes numbers can be organized and displayed in different ways, including in a table and in a scatter plot. A **scatter plot** is a set of disconnected data points plotted on a coordinate plane.

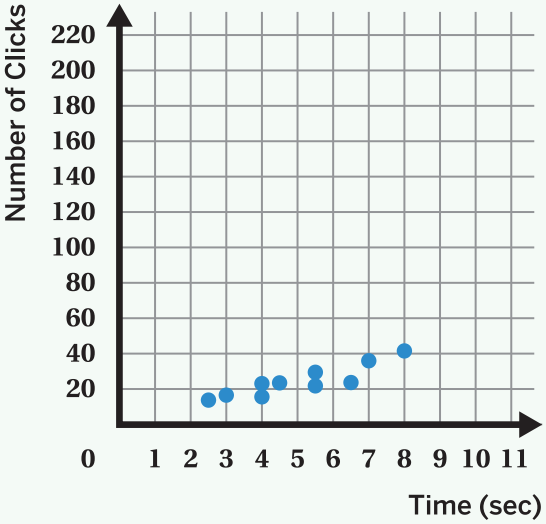
A table and a scatter plot both display the same data, but can be helpful in different ways. A scatter plot can be used to investigate connections between two variables, while a table is helpful for looking for the exact values of specific data points.

Here is data showing the amount of time in seconds and the number of clicks of the button.

**Table**

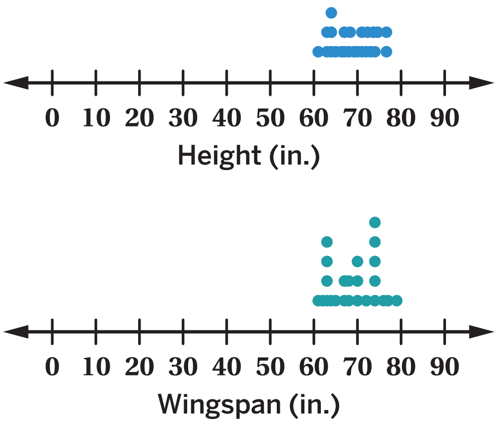
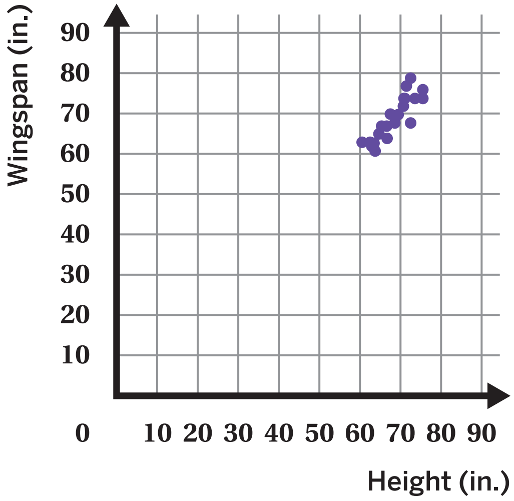
| Time (sec) | Number of Clicks |
|---|---|
| 2.5 | 14 |
| 3 | 17 |
| 4 | 16 |
| 4 | 23 |
| 4.5 | 24 |
| 5.5 | 22 |
| 5.5 | 30 |
| 6.5 | 24 |
| 7 | 36 |
| 8 | 42 |

**Scatter Plot**



## 8 Synthesis

Here is the height and wingspan data from a different class. Select one of the representations of the data and describe its advantages.

| Dot Plot | Scatter Plot |
|---|---|
|  |  |

## Summary

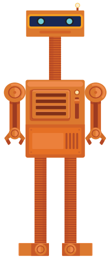
Data presented as numbers, quantities, or measurements that can be compared in a meaningful way is called *numerical data*, or *quantitative data*. In this lesson, we investigated *univariate data*, which involves one variable, and *bivariate data*, which involves two variables.

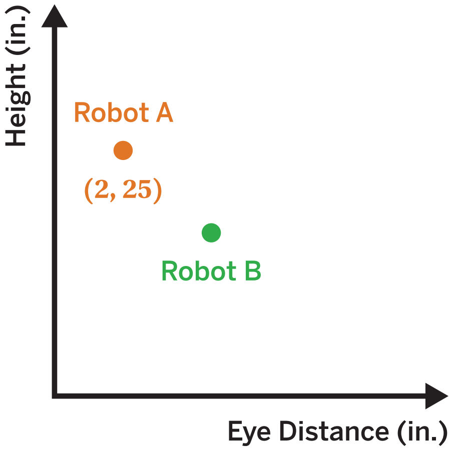
There are different ways to represent numerical data. A dot plot shows data for one variable and a scatter plot shows data for two variables at the same time. Seeing two numerical variables at the same time allows us to notice trends and connections.

## 7 Synthesis

This graph shows the height and eye distance for two robots.

Describe some things you know about Robot B given the information about Robot A.

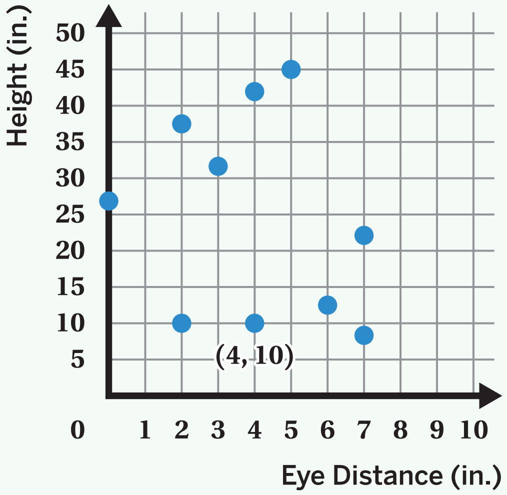




## Summary

A point on a *scatter plot* represents two pieces of information. The axis labels tell you how to interpret the coordinates of each point.

In this example, the point $(4, 10)$ represents a robot with an eye distance of $4$ inches and a height of $10$ inches.

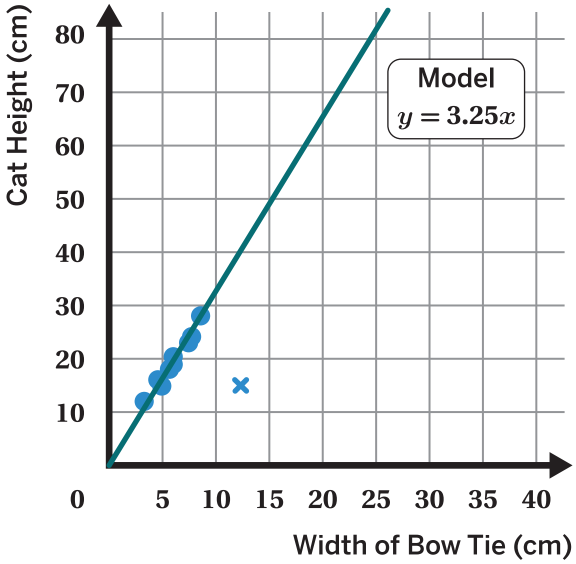


## 13 Synthesis

Discuss both questions. Then select one and write your response.

A linear model can be represented as an equation or a line. Why is a linear model helpful?

How can you identify an outlier on a scatter plot?

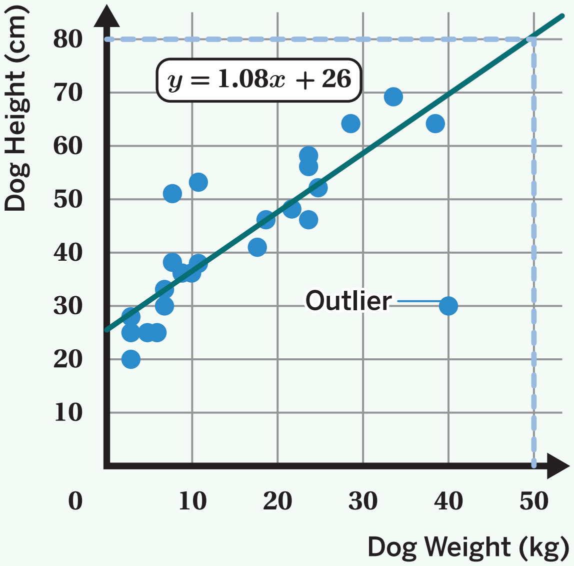


## Summary

A **linear model** is a line on a scatter plot that helps identify trends in data more clearly. It can also be used to make a prediction.

For example, there are two ways you can use a linear model to predict a dog's height when it weighs 50 kilograms.

- Use the graph to locate 50 on the $x$-axis and follow it up to meet the linear model, which shows a $y$-value of 80. This means when the dog's weight is 50 kilograms, its height is 80 centimeters.
- Use the equation for the linear model, $y = 1.08x + 26$, by replacing $x$ with 50 and evaluating for $y$, which is approximately 80 centimeters.



You can identify an **outlier** by looking for points that are far away from the other points and from the predicted values. The point (40, 30) is an outlier on the graph of dog weights and heights.

## Synthesis

How can a scatter plot help make sense of the relationship between two variables? Use the scatter plot you created in Activity 1 to support your thinking.

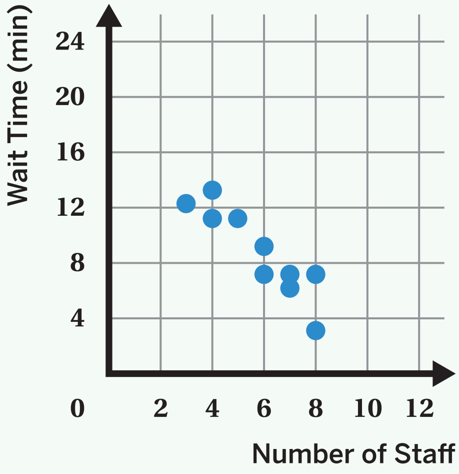
## Summary

Scatter plots show data points, patterns in data, and relationships between two variables in different ways.

For example, this scatter plot shows data about how long customers waited at a drive-thru restaurant and the number of staff working at that time.

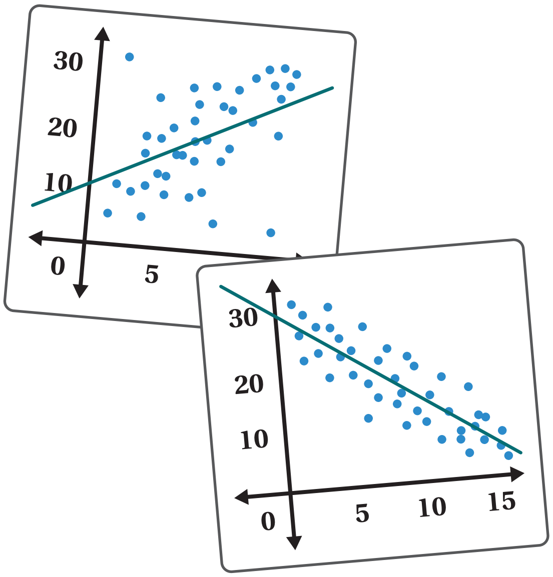
The scatter plot shows both specific information, as well as general trends, including:

- When 3 staff were working, the wait time was about 12 minutes.
- The more staff there are, the shorter the wait time seems to be.



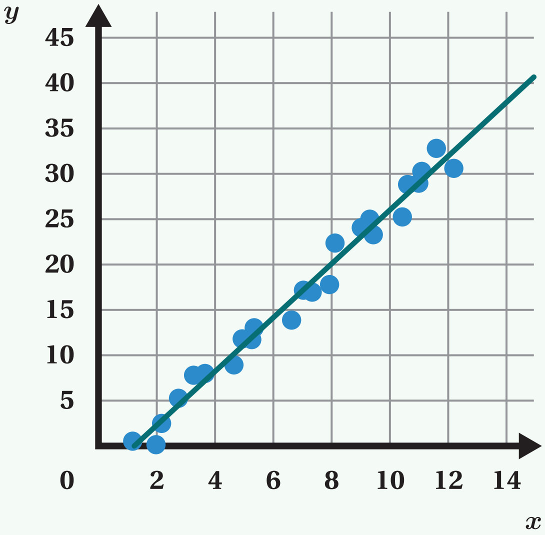
## 8 Synthesis

What are some things to consider when creating a line of fit? Use the examples if they help with your thinking.

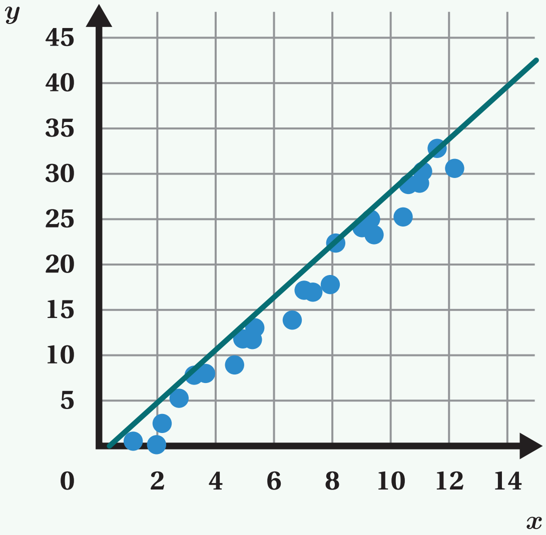


## Summary

When creating a line of fit for a scatter plot, it's important to determine how well the line fits the data. A good line of fit follows the trend of the data, is as close to the plotted points as possible, and has about the same number of points above and below the line. The line may pass through some, all, or none of the points.



This line is a good fit for the data.

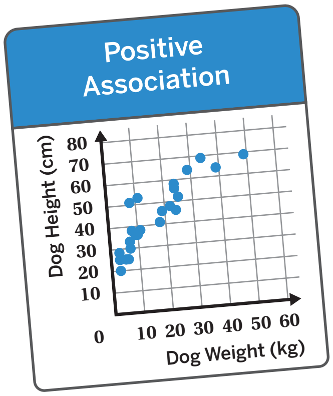


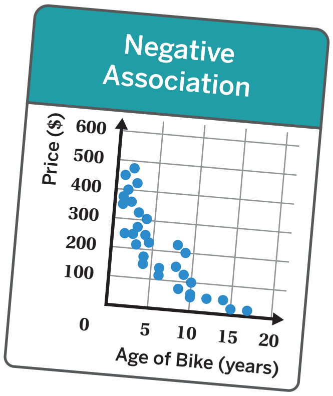
This line is not a good fit for the data.

## 11 Synthesis

Discuss both questions. Choose *one* and write your response.

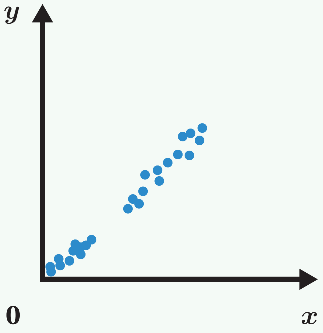
- What are some clues that a scatter plot might have a positive or negative association?
- What does the slope of a linear model tell you about the data?



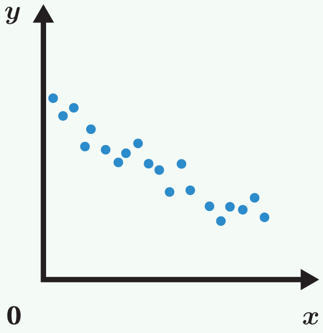


## Summary

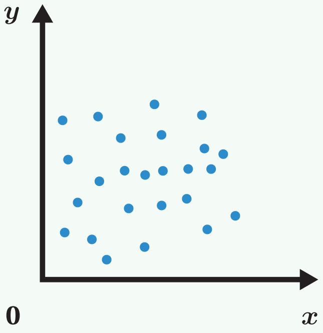
When two variables on a scatter plot are related, we call that an association. The slope of a linear model can help determine the type of association. A **positive association** means that when one variable increases, the other also increases. A **negative association** means that when one variable increases, the other decreases. If the scatter plot shows no clear trend between the two variables, then the variables have no association.



**Positive association**



**Negative association**



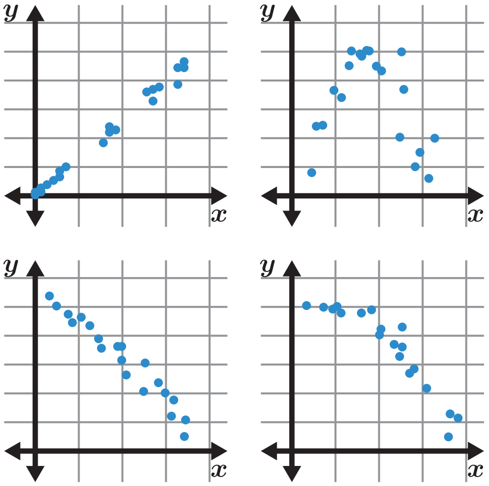
**No association**

## 7 Synthesis

Discuss both questions. Then select *one* and record your response.

- What is a non-linear association?

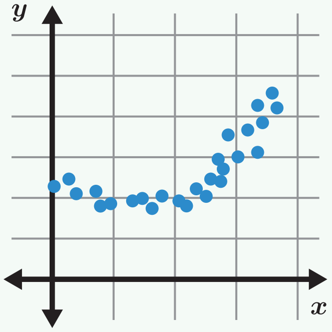
- What does it mean if a scatter plot has clusters?



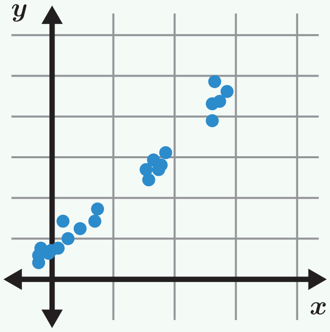
## Summary

When data on a scatter plot can be modeled with a straight line, we say it has a **linear association**. Data that can't be modeled by a straight line has a **non-linear association**. Sometimes groups of data points appear close together, which are called **clusters**.

This scatter plot is an example of a non-linear association, without clusters.



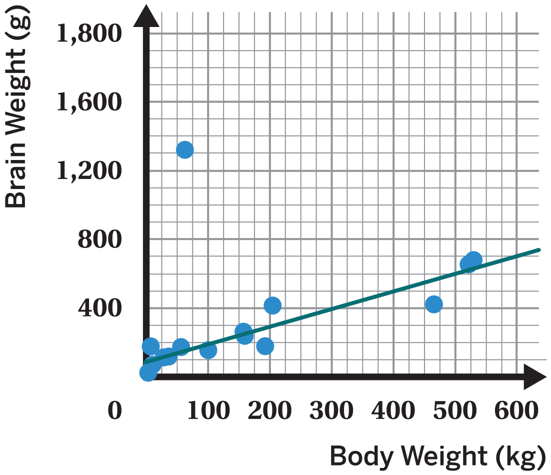
This scatter plot is an example of a linear association, with clusters.



## 11 Synthesis

Describe an advantage and disadvantage of using a line of fit to make predictions.

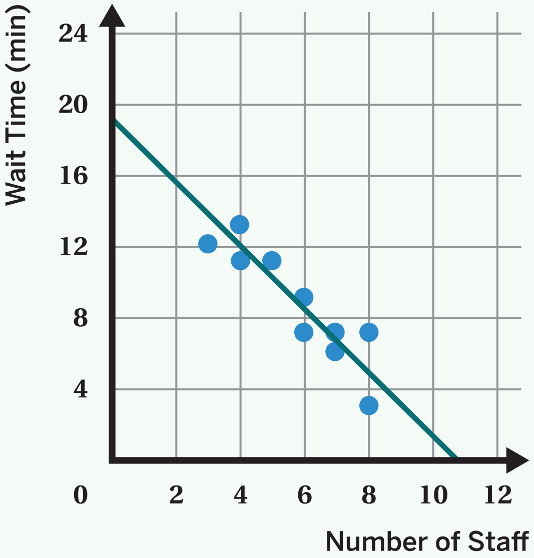
Use the graph if it helps with your thinking.



## Summary

By understanding the association between two variables, we can make predictions about unknown values. When there's a linear association, using a linear model can often make predictions more accurate.

For example, this scatter plot shows data about how many minutes customers waited at a drive-through restaurant and the number of staff working at that time. This data can be modeled by the equation $y = -1.75x + 19$.

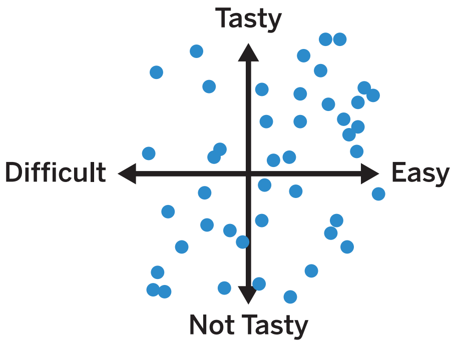


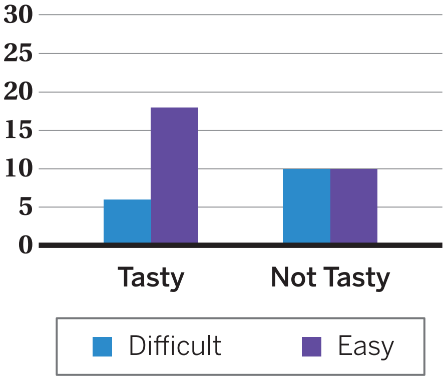
- The *slope* of the linear model is $-1.75$, which means that if the number of staff increases by $1$ person, the wait time decreases by $1.75$ minutes.
- The linear model predicts that if there are $2$ staff working, the wait time will be approximately $15.5$ minutes.
- But the linear model also predicts that when there are $0$ staff working, the wait time will be $19$ minutes, which is impossible!

## 10 Synthesis

A school surveyed the 8th graders about how tasty bananas are, and how easy they are to eat.

| | Difficult | Easy | Total |
|---|---|---|---|
| Tasty | 6 | 18 | 24 |
| Not Tasty | 10 | 10 | 20 |
| Total | 16 | 28 | 44 |





Select one representation for the data and describe its advantages.

## Summary

A **two-way table** lets us compare two variables of *categorical data*, which is data that can be sorted into categories. Two-way tables show one of the variables across the top and the other down one side. Each entry in the table represents the **frequency**, or the number of times, that a category appears in the data set.

For example, this two-way table shows data about whether students meditated on a certain day, and whether they felt calm or agitated that day.

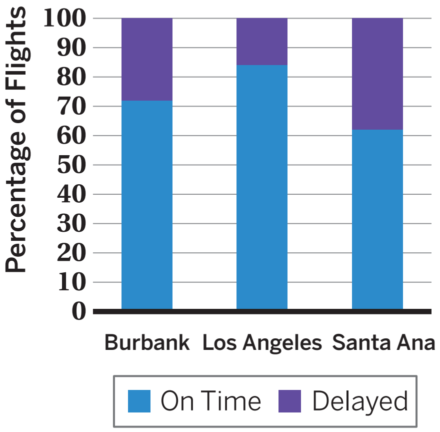
| | Meditated | Did Not Meditate | Total |
|---|---|---|---|
| Calm | 45 | 8 | 53 |
| Agitated | 23 | 21 | 44 |
| Total | 68 | 29 | 97 |

Two-way tables, scatter plots, and bar graphs can all be used to represent data and explore associations within data. Each representation has advantages and disadvantages. You can use these representations to investigate possible connections between variables. In the example, we can see there's a connection between meditating and feeling calm, since a majority of the people who felt calm also meditated.

## 12 Synthesis

Here are two representations of relative frequencies.

| | On Time | Delayed | Total |
|---|---|---|---|
| Burbank | 72% | 28% | 100% |
| L.A. | 84% | 16% | 100% |
| Santa Ana | 62% | 38% | 100% |



How can you use relative frequencies to identify possible associations between variables?

## Summary

We can use specific types of two-way tables and bar graphs to show frequencies and percentages within data sets.

The **relative frequency** of a category is the fraction or percentage of the data set that's in that category. A two-way table of relative frequencies shows the fraction or percentage of each category instead of the number of data points.

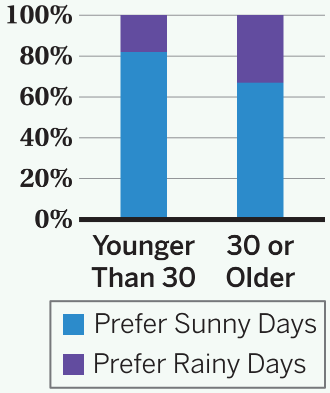
**Relative Frequencies**

| | Prefer Texting to Communicate | Prefer Making a Phone Call | Total |
|---|---|---|---|
| Younger Than 40 | 82% | 18% | 100% |
| 40 or Older | 33% | 67% | 100% |

A **segmented bar graph** compares different categories within a data set. Each bar represents all the data within one category, or 100%. The bars are each separated into parts, or segments, that show what percentage each part makes up of the whole category.

We can use representations like these to identify associations between two categorical variables. Categorical variables represent data that can be broken down into groups. For example, the table and graph below show an association between categorical variables, age, and communication preference.

**Segmented Bar Graph**


Younger Than 30
30 or Older
🟦 Prefer Sunny Days
🟪 Prefer Rainy Days